Rec'd PCT/PTO 14 FEB 2005

10/524355

CELL ADHESION AND EXTRACELLULAR MATRIX PROTEINS

TECHNICAL FIELD

The invention relates to novel nucleic acids, cell adhesion and extracellular matrix proteins encoded by these nucleic acids, and to the use of these nucleic acids and proteins in the diagnosis, treatment, and prevention of immune system disorders, neurological disorders, developmental disorders, connective tissue disorders, and cell proliferative disorders, including cancer. The invention also relates to the assessment of the effects of exogenous compounds on the expression of nucleic acids and cell adhesion and extracellular matrix proteins.

BACKGROUND OF THE INVENTION

Cell Adhesion Proteins

The surface of a cell is rich in transmembrane proteoglycans, glycoproteins, glycolipids, and receptors. These macromolecules mediate adhesion with other cells and with components of the ECM. The interaction of the cell with its surroundings profoundly influences cell shape, strength, flexibility, motility, and adhesion. These dynamic properties are intimately associated with signal transduction pathways controlling cell proliferation and differentiation, tissue construction, and embryonic development. Families of cell adhesion molecules include the cadherins, integrins, lectins, neural cell adhesion proteins, and some members of the proline-rich proteins.

Cadherins comprise a family of calcium-dependent glycoproteins that function in mediating cell-cell adhesion in virtually all solid tissues of multicellular organisms. These proteins share multiple repeats of a cadherin-specific motif, and the repeats form the folding units of the cadherin extracellular domain. Cadherin molecules cooperate to form focal contacts, or adhesion plaques, between adjacent epithelial cells. The cadherin family includes the classical cadherins and protocadherins. Classical cadherins include the E-cadherin, N-cadherin, and P-cadherin subfamilies. E-cadherin is present on many types of epithelial cells and is especially important for embryonic development. N-cadherin is present on nerve, muscle, and lens cells and is also critical for embryonic development. P-cadherin is present on cells of the placenta and epidermis. Recent studies report that protocadherins are involved in a variety of cell-cell interactions (Suzuki, S.T. (1996) J. Cell Sci. 109:2609-2611). The intracellular anchorage of cadherins is regulated by their dynamic association with catenins, a family of cytoplasmic signal transduction proteins associated with the actin cytoskeleton. The anchorage of cadherins to the actin cytoskeleton appears to be regulated by protein tyrosine phosphorylation, and the cadherins are the target of phosphorylation-induced junctional

disassembly (Aberle, H. et al. (1996) J. Cell. Biochem. 61:514-523).

Integrins are ubiquitous transmembrane adhesion molecules that link the ECM to the internal cytoskeleton. Integrins are composed of two noncovalently associated transmembrane glycoprotein subunits called α and β. At least 8 different ß subunits (ß1-ß8) and at least 12 different α subunits have been identified (α1-α8, αL, αM, αX, and αIIb). Individual α subunits are capable of associating with different ß subunits, suggesting a possible mechanism for specifying integrin function and ligand binding affinity. Members of the ß subunit family are generally of 90-110 kilodaltons (kD) in molecular weight and share about 40-48% amino acid sequence homology. About 56 cysteines distributed among four repeating units are also conserved. Some variation in these conserved features is observed among some of the more divergent ß subunit family members. Members of the α subunit family are generally 150-200 kilodaltons in molecular weight and are not as well conserved as the ß subunit family. All contain seven repeating domains of 24-45 amino acids spaced about 20-35 amino acids apart. The N-termini each contain 3-4 divalent cation binding sites. (For review, see Pigott, R. and C. Power (1994) The Adhesion Molecule Facts Book, Academic Press, San Diego, CA, pp. 9-12.)

Integrins function as receptors that specifically recognize and bind to ECM proteins such as fibronectin, fibrinogen, laminin, thrombospondin, vitronectin, von Willebrand factor, and collagen. Some integrins recognize a specific motif, the RGD sequence, at the C-termini of the ECM proteins they bind. Integrins also bind to immunoglobulin superfamily proteins such as ICAM-1, -2, and -3 and VCAM-1.

Most integrins have been shown to activate focal adhesion kinase (FAK), a protein tyrosine kinase that is linked to Ras signaling pathways that modify the cytoskeleton and stimulate the mitogen-activated protein kinase (MAPK) cascade (Hanks, S.K. and T.R. Polte (1997) BioEssays 19:137-145). Integrins can also influence growth factor signaling through direct interaction with growth factor receptor tyrosine kinases (RTKs) (Miyamoto, S. et al. (1996) J. Cell Biol. 135:1633-1642). Integrins have also been shown to play a vital role in "anoikis," a term describing programmed cell death caused by loss of cell anchorage (Frisch, S.M. and E. Ruoslahti (1997) Curr. Opin. Cell Biol. 9:701-706).

A number of diseases have been attributed to integrin defects. (See Pigott and Power, *supra*). For example, leukocyte adhesion deficiency (LAD) is an inherited disorder characterized by the impaired migration of neutrophils to sites of extravascular inflammation. LAD is caused by abnormal splicing of and a missense mutation in the RNA encoding the ß2 subunit. Additionally, defects in platelet integrin are correlated with Glanzmann's thrombasthemia, a bleeding disorder characterized by insufficient platelet aggregation.

Lectins comprise a ubiquitous family of extracellular glycoproteins which bind cell surface carbohydrates specifically and reversibly, resulting in the agglutination of cells (reviewed in Drickamer, K. and M.E. Taylor (1993) Annu. Rev. Cell Biol. 9:237-264). This function is particularly important for activation of the immune response. Lectins mediate the agglutination and mitogenic stimulation of lymphocytes at sites of inflammation (Lasky, L.A. (1991) J. Cell. Biochem. 45:139-146; Paietta, E. et al. (1989) J. Immunol. 143:2850-2857).

Lectins are further classified into subfamilies based on carbohydrate-binding specificity and other criteria. The galectin subfamily, in particular, includes lectins that bind β-galactoside carbohydrate moieties in a thiol-dependent manner (reviewed in Hadari, Y.R. et al. (1998) J. Biol. Chem. 270:3447-3453). Galectins are widely expressed and developmentally regulated. Galectins contain a characteristic carbohydrate recognition domain (CRD). The CRD comprises about 140 amino acids and contains several stretches of about 1 - 10 amino acids which are highly conserved among all galectins. A particular 6-amino acid motif within the CRD contains conserved tryptophan and arginine residues which are critical for carbohydrate binding. The CRD of some galectins also contains cysteine residues which may be important for disulfide bond formation. Secondary structure predictions indicate that the CRD forms several β-sheets.

Galectins play a number of roles in diseases and conditions associated with cell-cell and cell-matrix interactions. For example, certain galectins associate with sites of inflammation and bind to cell surface immunoglobulin E molecules. In addition, galectins may play an important role in cancer metastasis. Galectin overexpression is correlated with the metastatic potential of cancers in humans and mice. Moreover, anti-galectin antibodies inhibit processes associated with cell transformation, such as cell aggregation and anchorage-independent growth (see, for example, Su, Z.-Z. et al. (1996) Proc. Natl. Acad. Sci. USA 93:7252-7257).

Selectins, or LEC-CAMs, comprise a specialized lectin subfamily involved primarily in inflammation and leukocyte adhesion (Reviewed in Lasky, *supra*). Selectins mediate the recruitment of leukocytes from the circulation to sites of acute inflammation and are expressed on the surface of vascular endothelial cells in response to cytokine signaling. Selectins bind to specific ligands on the leukocyte cell membrane and enable the leukocyte to adhere to and migrate along the endothelial surface. Binding of selectin to its ligand leads to polarized rearrangement of the actin cytoskeleton and stimulates signal transduction within the leukocyte (Brenner, B. et al. (1997) Biochem. Biophys. Res. Commun. 231:802-807; Hidari, K. I. et al. (1997) J. Biol. Chem. 272:28750-28756). Members of the selectin family possess three characteristic motifs: a lectin or carbohydrate recognition domain; an epidermal growth factor-like domain; and a variable number of short consensus repeats (scr or "sushi"

repeats) which are also present in complement regulatory proteins.

Neural cell adhesion proteins (NCAPs) play roles in the establishment of neural networks during development and regeneration of the nervous system (Uyemura, K. et al. (1996) Essays Biochem. 31:37-48; Brummendorf, T., and F.G. Rathjen (1996) Curr. Opin. Neurobiol. 6:584-593). NCAP participates in neuronal cell migration, cell adhesion, neurite outgrowth, axonal fasciculation, pathfinding, synaptic target-recognition, synaptic formation, myelination and regeneration. NCAPs are expressed on the surfaces of neurons associated with learning and memory. Mutations in genes encoding NCAPS are linked with neurological diseases, including hereditary neuropathy, Charcot-Marie-Tooth disease, Dejerine-Sottas disease, X-linked hydrocephalus, MASA syndrome (mental retardation, aphasia, shuffling gait and adducted thumbs), and spastic paraplegia type I. In some cases, expression of NCAP is not restricted to the nervous system. L1, for example, is expressed in melanoma cells and hematopoietic tumor cells where it is implicated in cell spreading and migration, and may play a role in tumor progression (Montgomery, A.M. et al. (1996) J. Cell Biol. 132:475-485).

NCAPs have at least one immunoglobulin constant or variable domain (Uyemura et al., *supra*). They are generally linked to the plasma membrane through a transmembrane domain and/or a glycosyl-phosphatidylinositol (GPI) anchor. The GPI linkage can be cleaved by GPI phospholipase C. Most NCAPs consist of an extracellular region made up of one or more immunoglobulin domains, a membrane spanning domain, and an intracellular region. Many NCAPs contain post-translational modifications including covalently attached oligosaccharide, glucuronic acid, and sulfate. NCAPs fall into three subgroups: simple-type, complex-type, and mixed-type. Simple-type NCAPs contain one or more variable or constant immunoglobulin domains, but lack other types of domains. Members of the simple-type subgroup include Schwann cell myelin protein (SMP), limbic system-associated membrane protein (LAMP), opiate-binding cell-adhesion molecule (OBCAM), and myelin-associated glycoprotein (MAG). The complex-type NCAPs contain fibronectin type III domains in addition to the immunoglobulin domains. The complex-type subgroup includes neural cell-adhesion molecule (NCAM), axonin-1, F11, Bravo, and L1. Mixed-type NCAPs contain a combination of immunoglobulin domains and other motifs such as tyrosine kinase and epidermal growth factor-like domains. This subgroup includes Trk receptors of nerve growth factors such as nerve growth factor (NGF) and neurotropin 4 (NT4), Neu differentiation factors such as glial growth factor II (GGFII) and acetylcholine receptor-inducing factor (ARIA), and the semaphorin/collapsin family such as semaphorin B and collapsin.

Semaphorins are a large group of axonal guidance molecules consisting of at least 30 different

members and are found in vertebrates, invertebrates, and even certain viruses. All semaphorins contain the sema domain which is approximately 500 amino acids in length. Neuropilin, a semaphorin receptor, has been shown to promote neurite outgrowth *in vitro*. The extracellular region of neuropilins consists of three different domains: CUB, discoidin, and MAM domains. The CUB and the MAM motifs of neuropilin have been proposed to have roles in protein-protein interactions and are suggested to be involved in the binding of semaphorins through the sema and the C-terminal domains (reviewed in Raper, J.A. (2000) Curr. Opin. Neurobiol. 10:88-94).

An NCAP subfamily, the NCAP-LON subgroup, includes cell adhesion proteins expressed on distinct subpopulations of brain neurons. Members of the NCAP-LON subgroup possess three immunoglobulin domains and bind to cell membranes through GPI anchors. Kilon (a kindred of NCAP-LON), for example, is expressed in the brain cerebral cortex and hippocampus (Funatsu, N. et al. (1999) J. Biol. Chem. 274:8224-8230). Immunostaining localizes Kilon to the dendrites and soma o pyramidal neurons. Kilon has three C2 type immunoglobulin-like domains, six predicted glycosylation sites, and a GPI anchor. Expression of Kilon is developmentally regulated. It is expressed at higher levels in adult brain in comparison to embryonic and early postnatal brains. Confocal microscopy shows the presence of Kilon in dendrites of hypothalamic magnocellular neurons secreting neuropeptides, oxytocin or arginine vasopressin (Miyata, S. et al. (2000) J. Comp. Neurol. 424:74-85). Arginine vasopressin regulates body fluid homeostasis, extracellular osmolarity and intravascular volume. Oxytocin induces contractions of uterine smooth muscle during child birth and of myoepithelial cells in mammary glands during lactation. In magnocellular neurons, Kilon is proposed t play roles in the reorganization of dendritic connections during neuropeptide secretion.

The co-ordinated function of effector and accessory cells in the immune system is assisted by adhesion molecules on the cell surface that stabilize interactions between different cell types. Leukocyte function-associated antigen 1 (LFA-1) is expressed on the surface of all white blood cells and is a receptor for intercellular adhesion molecules (ICAM) 1 and 2 which are members of the immunoglobulin superfamily. The interaction of LFA-1 with ICAMs 1 and 2 provides essential accessory adhesion signals in many immune interactions, including those between T and B lymphocytes and cytotoxic T cells and their targets. In addition, both ICAMs are expressed at low levels on resting vascular endothelium. ICAM-1 is strongly upregulated by cytokine stimulation and plays a key role in the arrest of leukocytes in blood vessels at sites of inflammation and injury. A third ligand for LFA-1 expressed in resting leukocytes is ICAM-3. ICAM-3 is closely related to ICAM-1 and is constitutively expressed on all leukocytes. It consists of five immunoglobulin domains and binds LFA-1 through its two N-terminal domains (Fawcett, J. et al. (1992) Nature 360:481-484).

Cell adhesion proteins also include some members of the proline-rich proteins (PRPs). PRPs are defined by a high frequency of proline, ranging from 20-50% of the total amino acid content. Some PRPs have short domains which are rich in proline. These proline-rich regions are associated with protein-protein interactions. One family of PRPs are the proline-rich synapse-associated proteins (ProSAPs) which have been shown to bind to members of the postsynaptic density (PSD) protein family and subtypes of the somatostatin receptor (Yao, I. et al. (1999) J. Biol. Chem. 274: 27463-27466; Zitzer, H. et al. (1999) J. Biol. Chem. 274:32997-33001). Members of the ProSAP family contain six to seven ankyrin repeats at the N-terminus, followed by an SH3 domain, a PDZ domain, and seven proline-rich regions and a SAM domain at the C terminus. Several groups of ProSAPs are important structural constituents of synaptic structures in human brain (Zitzer et al., *supra*). Another member of the PRP family is the HLA-B-associated transcript 2 protein (BAT2) which is rich in proline and includes short tracts of polyproline, polyglycine, and charged amino acids. BAT2 also contains four RGD (Arg-Gly-Asp) motifs typical of integrins (Banerji, J. et al. (1990) Proc. Natl. Acad. Sci. USA 87:2374-2378).

Toposome is a cell-adhesion glycoprotein isolated from mesenchyme-blastula embryos. Toposome precursors including vitellogenin promote cell adhesion of dissociated blastula cells.

There are additional specific domains characteristic of cell adhesion proteins. One such domain is the MAM domain, a domain of about 170 amino acids found in the extracellular region of diverse proteins. These proteins all share a receptor-like architecture comprising a signal peptide, followed by a large N-terminal extracellular domain, a transmembrane region, and an intracellular domain (PROSITE document PDOC00604 MAM domain signature and profile). MAM domain proteins include zonadhesin, a sperm-specific membrane protein that binds to the zona pellucida of the egg; neuropilin, a cell adhesion molecule that functions during the formation of certain neuronal circuits, and *Xenopus laevis* thyroid hormone induced protein B, which contains four MAM domains and is involved in metamorphosis (Brown, D.D. et al. (1996) Proc. Natl. Acad. Sci. USA 93:1924-1929).

The WSC domain was originally found in the yeast WSC (cell-wall integrity and stress response component) proteins which act as sensors of environmental stress. The WSC domains are extracellular and are thought to possess a carbohydrate binding role (Ponting, C.P. et al. (1999) Curr. Biol. 9:S1-S2). A WSC domain has recently been identified in polycystin-1, a human plasma membrane protein. Mutations in polycystin-1 are the cause of the commonest form of autosomal dominant polycystic kidney disease (Ponting, C.P. et al. (1999) Curr. Biol. 9:R585-R588).

Leucine rich repeats (LRR) are short motifs found in numerous proteins from a wide range of

species. LRR motifs are of variable length, most commonly 20-29 amino acids, and multiple repeats are typically present in tandem. LRR motifs are important for protein/protein interactions and cell adhesion, and LRR proteins are involved in cell/cell interactions, morphogenesis, and development (Kobe, B. and J. Deisenhofer (1995) Curr. Opin. Struct. Biol. 5:409-416). The human ISLR (immunoglobulin superfamily containing leucine-rich repeat) protein contains a C2-type immunoglobulin domain as well as LRR motifs. The ISLR gene is linked to the critical region for Bardet-Biedl syndrome, a developmental disorder of which the most common feature is retinal dystrophy (Nagasawa, A. et al. (1999) Genomics 61:37-43).

The sterile alpha motif (SAM) domain is a conserved protein binding domain, approximately 70 amino acids in length, and is involved in the regulation of many developmental processes in eukaryotes. The SAM domain can potentially function as a protein interaction module through its ability to form homo- or hetero-oligomers with other SAM domains (Schultz, J. et al. (1997) Protein Sci. 6:249-253).

Vinculin is a cellular adhesion molecule that is involved in the attachment of actin microfilaments to the plasma membrane of eukaryotic cells. This protein is composed of approximately 1000 amino acid residues and is characterized by an acidic N-terminal domain consisting of either two (in C. elegans) or three (in vertebrates) repeats of a 110 amino acid region. A proline-rich region is followed by a basic C-terminal domain. Two signature patterns are found in the N-terminal domain, one which seems to be involved in protein-protein interactions and one based on the repeated region (PROSITE document PDOC00568 Vinculin family signatures).

Synapsins are a family of proteins that coat synaptic vesicles and bind to actin filaments as well as other components of the cytoskeleton. Synapsins I and II each exist as two alternately spliced variants termed IA and IB or IIA and IIB and differ from each other in their C-termini. Two conserved domains among these proteins are an octapeptide consisting of a phosphorylated serine residue and a second domain of a stretch of 11 highly conserved residues (PROSITE document PDOC00345 Synapsin signatures).

Osteonectin domain signatures are derived from three extracellular proteins (SPARC or osteonectin, SC1, and QR1) which contain a region of 240 highly-conserved amino acid residues in their C-termini. Two signature patterns were developed based on this conserved region, one based on a cysteine-rich region and the other based on a stretch of 11 highly conserved residues (PROSITE document PDOC00535 Osteonectin domain signatures).

Extracellular Matrix Proteins

The extracellular matrix (ECM) is a complex network of glycoproteins, polysaccharides,

proteoglycans, and other macromolecules that are secreted from the cell into the extracellular space. The ECM remains in close association with the cell surface and provides a supportive meshwork that profoundly influences cell shape, motility, strength, flexibility, and adhesion. In fact, adhesion of a cell to its surrounding matrix is required for cell survival except in the case of metastatic tumor cells, which have overcome the need for cell-ECM anchorage. This phenomenon suggests that the ECM plays a critical role in the molecular mechanisms of growth control and metastasis. (Reviewed in Ruoslahti, E. (1996) Sci. Am. 275:72-77.) Furthermore, the ECM determines the structure and physical properties of connective tissue and is particularly important for morphogenesis and other processes associated with embryonic development and pattern formation.

The collagens comprise a family of ECM proteins that provide structure to bone, teeth, skin, ligaments, tendons, cartilage, blood vessels, and basement membranes. Multiple collagen proteins have been identified. Three collagen molecules fold together in a triple helix stabilized by interchain disulfide bonds. Bundles of these triple helices then associate to form fibrils. Collagen primary structure consists of hundreds of (Gly-X-Y) repeats where about a third of the X and Y residues are Pro. Glycines are crucial to helix formation as the bulkier amino acid sidechains cannot fold into the triple helical conformation. Because of these strict sequence requirements, mutations in collagen genes have severe consequences. Osteogenesis imperfecta patients have brittle bones that fracture easily; in severe cases patients die *in utero* or at birth. Ehlers-Danlos syndrome patients have hyperelastic skin, hypermobile joints, and susceptibility to aortic and intestinal rupture. Chondrodysplasia patients have short stature and ocular disorders. Alport syndrome patients have hematuria, sensorineural deafness, and eye lens deformation. (Isselbacher, K.J. et al. (1994) Harrison's Principles of Internal Medicine, McGraw-Hill, Inc., New York, NY, pp. 2105-2117; and Creighton, T.E. (1984) Proteins, Structures and Molecular Principles, W.H. Freeman and Company, New York, NY, pp. 191-197.)

Elastin and related proteins confer elasticity to tissues such as skin, blood vessels, and lungs. Elastin is a highly hydrophobic protein of about 750 amino acids that is rich in proline and glycine residues. Elastin molecules are highly cross-linked, forming an extensive extracellular network of fibers and sheets. Elastin fibers are surrounded by a sheath of microfibrils which are composed of a number of glycoproteins, including fibrillin. Mutations in the gene encoding fibrillin are responsible for Marfan's syndrome, a genetic disorder characterized by defects in connective tissue. In severe cases, the aortas of afflicted individuals are prone to rupture. (Reviewed in Alberts, B. et al. (1994) Molecular Biology of the Cell, Garland Publishing, New York, NY, pp. 984-986.) The fibulin proteins connect elastic fibers and are though to promote the formation and stabilization of the fiber. Members

of the fibulin family contain epidermal growth factor-like motifs as well as an RGD cell attachment sequence (Midwood, K.S. and J.E. Schwarzbauer (2002) Current Biology 12:R279-R281).

Fibronectin is a large ECM glycoprotein found in all vertebrates. Fibronectin exists as a dimer of two subunits, each containing about 2,500 amino acids. Each subunit folds into a rod-like structure containing multiple domains. The domains each contain multiple repeated modules, the most common of which is the type III fibronectin repeat. The type III fibronectin repeat is about 90 amino acids in length and is also found in other ECM proteins and in some plasma membrane and cytoplasmic proteins. Furthermore, some type III fibronectin repeats contain a characteristic tripeptide consisting of Arginine-Glycine-Aspartic acid (RGD). The RGD sequence is recognized by the integrin family of cell surface receptors and is also found in other ECM proteins. Disruption of both copies of the gene encoding fibronectin causes early embryonic lethality in mice. The mutant embryos display extensive morphological defects, including defects in the formation of the notochord, somites, heart, blood vessels, neural tube, and extraembryonic structures. (Reviewed in Alberts et al., *supra*, pp. 986-987.)

Laminin is a major glycoprotein component of the basal lamina which underlies and supports epithelial cell sheets. Laminin is one of the first ECM proteins synthesized in the developing embryo. Laminin is an 850 kilodalton protein composed of three polypeptide chains joined in the shape of a cross by disulfide bonds. Laminin is especially important for angiogenesis and, in particular, for guidir the formation of capillaries. (Reviewed in Alberts et al., *supra*, pp. 990-991.)

Many proteinaceous ECM components are proteoglycans. Proteoglycans are composed of unbranched polysaccharide chains (glycosaminoglycans) attached to protein cores. Common proteoglycans include aggrecan, betaglycan, decorin, perlecan, serglycin, and syndecan-1. Some of these molecules not only provide mechanical support, but also bind to extracellular signaling molecules such as fibroblast growth factor and transforming growth factor β, suggesting a role for proteoglycans in cell-cell communication. (Reviewed in Alberts et al., *supra*, pp. 973-978.) Likewise, the glycoproteins tenascin-C and tenascin-R are expressed in developing and lesioned neural tissue and provide stimulatory and anti-adhesive (inhibitory) properties, respectively, for axonal growth (Faissner, A. (1997) Cell Tissue Res. 290:331-341).

Dentin phosphoryn (DPP) is a major component of the dentin ECM. DPP is a proteoglycan that is synthesized and expressed by odontoblasts (Gu, K. et al. (1998) Eur. J. Oral Sci. 106:1043-1047). DPP is believed to nucleate or modulate the formation of hydroxyapatite crystals.

Amelogenin is an extracellular matrix protein that plays a role in the biomineralization in tooth enamel. This protein participates in the regulation of crystallite formation during tooth enamel development and thus, is thought to play a major role the structural organization and mineralization of

developing tooth enamel (Li, W. et al. (2001) Matrix Biol. 19(8):755-60).

Mucins are highly glycosylated glycoproteins that are the major structural component of the mucus gel. The physiological functions of mucins are cytoprotection, mechanical protection, maintenance of viscosity in secretions, and cellular recognition. MUC6 is a human gastric mucin that is also found in gall bladder, pancreas, seminal vesicles, and female reproductive tract (Toribara, N.W. et al. (1997) J. Biol. Chem. 272:16398-16403). The MUC6 gene has been mapped to human chromosome 11 (Toribara, N.W. et al. (1993) J. Biol. Chem. 268:5879-5885). Hemomucin is a novel *Drosophila* surface mucin that may be involved in the induction of antibacterial effector molecules (Theopold, U. et al. (1996) J. Biol. Chem. 217:12708-12715).

Olfactomedin was originally identified as the major component of the mucus layer surrounding the chemosensory dendrites of olfactory neurons. Olfactomedin-related proteins are secreted glycoproteins with conserved C-terminal motifs. The TIGR/myocilin protein, an olfactomedin-related protein expressed in the eye, is associated with the pathogenesis of glaucoma (Kulkarni, N.H. et al. (2000) Genet. Res. 76:41-50).

Ankyrin (ANK) repeats mediate protein-protein interactions associated with diverse intracellular functions. ANK repeats are composed of about 33 amino acids that form a helix-turn-helix core preceded by a protruding "tip." These tips are of variable sequence and may play a role in protein-protein interactions. The helix-turn-helix region of the ANK repeats stack on top of one another and are stabilized by hydrophobic interactions (Yang, Y. et al. (1998) Structure 6:619-626).

Sushi repeats, also called short consensus repeats (SCR), are found in a number of proteins that share the common feature of binding to other proteins. For example, in the C-terminal domain of versican, the sushi domain is important for heparin binding. Sushi domains contain basic amino acid residues, which may play a role in binding (Oleszewski, M. et al. (2000) J. Biol. Chem. 275:34478-34485).

Link, or X-link, modules are hyaluronan-binding domains found in proteins involved in the assembly of extracellular matrix, cell adhesion, and migration. The Link module superfamily includes CD44, cartilage link protein, and aggrecan. This family also includes BEHAB (brain enriched hyaluronan-binding)/brevican, a component of the brain ECM that is dramatically upregulated in human gliomas, and appears to play a role in determining the invasive potential of brain tumor cells (Gary, S.C. et al. (1998) Curr. Opin. Neurobiol. 8:576-581). There is close similarity between the Link module and the C-type lectin domain, with the predicted hyaluronan-binding site at an analogous position to the carbohydrate-binding pocket in E-selectin (Kohda, D. et al. (1996) Cell 86:767-775).

Multidomain or mosaic proteins play an important role in the diverse functions of the

extracellular matrix (Engel, J. et al. (1994) Development (Camb.):S35-S42). ECM proteins are frequently characterized by the presence of one or more domains which may contain a number of potential intracellular disulfide bridge motifs. For example, domains which match the epidermal growth factor (EGF) tandem repeat consensus are present within several known extracellular proteins that promote cell growth,, development, and cell signaling. This signature sequence is about forty amino acid residues in length and includes six conserved cysteine residues, and a calcium-binding site near the N-terminus of the signature sequence. The main structure is a two-stranded beta-sheet followed by a loop to a C-terminal short two-stranded sheet. Subdomains between the conserved cysteines vary in length (Davis, C.G. (1990) New Biol. 5:410-419). Post-translational hydroxylation of aspartic acid or asparagine residues has been associated with EGF-like domains in several proteins (Prosite PDOC00010).

A number of proteins that contain calcium-binding EGF-like domain signature sequences are involved in growth and differentiation. Examples include bone morphogenic protein 1, which induces the formation of cartilage and bone; crumbs, which is a *Drosophila* epithelial development protein; Notch and a number of its homologs, which are involved in neural growth and differentiation, and transforming growth factor beta-1 binding protein (Expasy PROSITE document PDOC00913; Soler, C. and G. Carpenter, in Nicola, N.A. (1994) The Cytokine Facts Book, Oxford University Press, Oxford, UK, pp. 193-197). EGF-like domains mediate protein-protein interactions for a variety of proteins. For example, EGF-like domains in the ECM glycoprotein fibulin-1 have been shown to mediate both self-association and binding to fibronectin (Tran, H. et al. (1997) J. Biol. Chem. 272:22600-22606). Point mutations in the EGF-like domains of ECM proteins have been identified as the cause of human disorders such as Marfan syndrome and pseudochondroplasia (Maurer, P. et al. (1996) Curr. Opin. Cell Biol. 8:609-617).

The CUB domain is an extracellular domain of approximately 110 amino acid residues found mostly in developmentally regulated proteins. The CUB domain contains four conserved cysteine residues and is predicted to have a structure similar to that of immunoglobulins. Vertebrate bone morphogenic protein 1, which induces cartilage and bone formation, and fibropellins I and III from sea urchin, which form the apical lamina component of the ECM, are examples of proteins that contain both CUB and EGF domains (PROSITE PDOC00908).

Other ECM proteins are members of the type A domain of von Willebrand factor (vWFA)-like module superfamily, a diverse group of proteins with a module sharing high sequence similarity. The vWFA-like module is found not only in plasma proteins but also in plasma membrane and ECM proteins (Colombatti, A. and P. Bonaldo (1991) Blood 77:2305-2315). Crystal structure analysis of an

integrin vWFA-like module shows a classic "Rossmann" fold and suggests a metal ion-dependent adhesion site for binding protein ligands (Lee, J.-O. et al. (1995) Cell 80:631-638). This family includes the protein matrilin-2, an extracellular matrix protein that is expressed in a broad range of mammalian tissues and organs. Matrilin-2 is thought to play a role in ECM assembly by bridging collagen fibrils and the aggrecan network (Deak, F. et al. (1997) J. Biol. Chem. 272:9268-9274).

The thrombospondins are multimeric, calcium-binding extracellular glycoproteins found widely in the embryonic extracellular matrix. These proteins are expressed in the developing nervous system or at specific sites in the adult nervous system after injury. Thrombospondins contain multiple EGF-type repeats, as well as a motif known as the thrombospondin type 1 repeat (TSR). The TSR is approximately 60 amino acids in length and contains six conserved cysteine residues. Motifs within TSR domains are involved in mediating cell adhesion through binding to proteoglycans and sulfated glycolipids. Thrombospondin-1 inhibits angiogenesis and modulates endothelial cell adhesion, motility, and growth. TSR domains are found in a diverse group of other proteins, most of which are expressed in the developing nervous system and have potential roles in the guidance of cell and growth cone migration. Proteins that contain TSRs include the F-spondin gene family, the semaphorin 5 family, UNC-5, and SCO-spondin. The TSR superfamily includes the ADAMTS proteins which contain an ADAM (A Disintegrin and Metalloproteinase) domain as well as one or more TSRs. The ADAMTS proteins have roles in regulating the turnover of cartilage matrix, regulation of blood vessel growth, and possibly development of the nervous system. (Reviewed in Adams, J.C. and R.P. Tucker (2000) Dev. Dyn. 218:280-299.)

Fibrinogen, the principle protein of vertebrate blood clotting, is a hexamer consisting of two sets of three different chains (alpha, beta, and gamma). The C-terminal domain of the beta and gamma chains comprises about 270 amino acid residues and contains four cysteines involved in two disulfide bonds. This domain has also been found in mammalian tenascin-X, an ECM protein that appears to be involved in cell adhesion (Prosite PDOC00445).

Expression profiling

Microarrays are analytical tools used in bioanalysis. A microarray has a plurality of molecules spatially distributed over, and stably associated with, the surface of a solid support. Microarrays of polypeptides, polynucleotides, and/or antibodies have been developed and find use in a variety of applications, such as gene sequencing, monitoring gene expression, gene mapping, bacterial identification, drug discovery, and combinatorial chemistry.

One area in particular in which microarrays find use is in gene expression analysis. Array technology can provide a simple way to explore the expression of a single polymorphic gene or the

expression profile of a large number of related or unrelated genes. When the expression of a single gene is examined, arrays are employed to detect the expression of a specific gene or its variants. When an expression profile is examined, arrays provide a platform for identifying genes that are tissue specific, are affected by a substance being tested in a toxicology assay, are part of a signaling cascade, carry out housekeeping functions, or are specifically related to a particular genetic predisposition, condition, disease, or disorder. The potential application of gene expression profiling is particularly relevant to improving diagnosis, prognosis, and treatment of disease. For example, both the levels and sequences expressed in tissues from subjects with diabetes may be compared with the levels and sequences expressed in normal tissue.

Jurkat Cells

Jurkat is an acute T cell leukemia cell line that grows actively in the absence of external stimuli. Jurkat has been extensively used to study signaling in human T cells. PMA (phorbol myristate acetate) is a broad activator of the protein kinase C-dependent pathways. Ionomycin is a calcium ionophore that permits the entry of calcium into the cell, hence increasing the cytosolic calcium concentration. The combination of PMA and ionomycin activates two of the major signaling pathways used by mammalian cells to interact with their environment. In T cells, the combination of PMA and ionomycin mimics the type of secondary signaling events elicited during optimal B cell activation.

Breast Cancer

More than 180,000 new cases of breast cancer are diagnosed each year, and the mortality rate for breast cancer approaches 10% of all deaths in females between the ages of 45-54 (Gish, K. (1999) AWIS Magazine 28:7-10). However the survival rate based on early diagnosis of localized breast cancer is extremely high (97%), compared with the advanced stage of the disease in which the tumor has spread beyond the breast (22%). Current procedures for clinical breast examination are lacking in sensitivity and specificity, and efforts are underway to develop comprehensive gene expression profiles for breast cancer that may be used in conjunction with conventional screening methods to improve diagnosis and prognosis of this disease (Perou, C.M. et al. (2000) Nature 406:747-752).

Mutations in two genes, BRCA1 and BRCA2, are known to greatly predispose a woman to breast cancer and may be passed on from parents to children (Gish, *supra*). However, this type of hereditary breast cancer accounts for only about 5% to 9% of breast cancers, while the vast majority of breast cancer is due to non-inherited mutations that occur in breast epithelial cells.

The relationship between expression of epidermal growth factor (EGF) and its receptor, EGFR, to human mammary carcinoma has been particularly well studied (Khazaie, K. et al. (1993)

Cancer and Metastasis Rev. 12:255-274, and references cited therein for a review of this area.) Overexpression of EGFR, particularly coupled with down-regulation of the estrogen receptor, is a marker of poor prognosis in breast cancer patients. In addition, EGFR expression in breast tumor metastases is frequently elevated relative to the primary tumor, suggesting that EGFR is involved in tumor progression and metastasis. This is supported by accumulating evidence that EGF has effects on cell functions related to metastatic potential, such as cell motility, chemotaxis, secretion and differentiation. Changes in expression of other members of the erbB receptor family, of which EGFR is one, have also been implicated in breast cancer. The abundance of erbB receptors, such as HER-2/neu, HER-3, and HER-4, and their ligands in breast cancer points to their functional importance in the pathogenesis of the disease, and may therefore provide targets for therapy of the disease (Bacus, S.S. et al. (1994) Am. J. Clin. Pathol. 102:S13-S24). Other known markers of breast cancer include a human secreted frizzled protein mRNA that is downregulated in breast tumors; the matrix G1a protein which is overexpressed in human breast carcinoma cells; Drg1 or RTP, a gene whose expression is diminished in colon, breast, and prostate tumors; maspin, a tumor suppressor gene downregulated in invasive breast carcinomas; and CaN19, a member of the S100 protein family, all of which are down-regulated in mammary carcinoma cells relative to normal mammary epithelial cells (Zhou, Z. et al. (1998) Int. J. Cancer 78:95-99; Chen, L. et al. (1990) Oncogene 5:1391-1395; Ulrix, W. et al (1999) FEBS Lett 455:23-26; Sager, R. et al. (1996) Curr. Top. Microbiol. Immunol. 213:51-64; and Lee, S.W. et al. (1992) Proc. Natl. Acad. Sci. USA 89:2504-2508).

Cell lines derived from human mammary epithelial cells at various stages of breast cancer provide a useful model to study the process of malignant transformation and tumor progression as it has been shown that these cell lines retain many of the properties of their parental tumors for lengthy culture periods (Wistuba, I.I. et al. (1998) Clin. Cancer Res. 4:2931-2938). Such a model is particularly useful for comparing phenotypic and molecular characteristics of human mammary epithelial cells at various stages of malignant transformation.

BT-20 is a breast carcinoma cell line derived *in vitro* from the cells emigrating out thin slices of the tumor mass isolated from a 74-year-old female. BT-474 is a breast ductal carcinoma cell line that was isolated from a solid, invasive ductal carcinoma of the breast obtained from a 60-year-old woman. BT-474 displays typical epithelial cellular structures such as desmosomes, microvilli, gap junctions, and tight junctions. This cell line has also discernable microtubules, tonofibrils, lysosomes, and osmiophilic secretory granules. BT-483 is a breast ductal carcinoma cell line that was isolated from a papillary invasive ductal tumor obtained from a 23-year-old normal, menstruating, parous female with a family history of breast cancer. BT-483 displays characteristic epithelial cellular

structures such as desmosomes, microvilli, tight junctions, and gap junctions. Hs 578T is a breast ductal carcinoma cell line that was isolated from a 74-year-old female with breast carcinoma. These cells do not express any detectable estrogen receptors and do not form colonies in semi-solid culture medium. MCF7 is a nonmalignant breast adenocarcinoma cell line isolated from the pleural effusion of a 69-year-old female. MCF7 has retained characteristics of the mammary epithelium such as the ability to process estradiol via cytoplasmic estrogen receptors and the capacity to form domes in culture. MCF-10A is a breast mammary gland (luminal ductal characteristics) cell line that was isolated from a 36-year-old woman with fibrocystic breast disease. MCF-10A expresses cytoplasmic keratins, epithelial sialomucins, and milkfat globule antigens. This cell lines exhibits three-dimensional growth in collagen and forms domes in confluent culture. MDA-MB-468 is breast adenocarcinoma cell line isolated from the pleural effusion of a 51-year-old female with metastatic adenocarcinoma of the breast.

Prostate Cancer

Prostate cancer is a common malignancy in men over the age of 50, and the incidence increases with age. In the US, there are approximately 132,000 newly diagnosed cases of prostate cancer and more than 33,000 deaths from the disorder each year.

Once cancer cells arise in the prostate, they are stimulated by testosterone to a more rapid growth. Thus, removal of the testes can indirectly reduce both rapid growth and metastasis of the cancer. Over 95 percent of prostatic cancers are adenocarcinomas which originate in the prostatic acini. The remaining 5 percent are divided between squamous cell and transitional cell carcinomas, both of which arise in the prostatic ducts or other parts of the prostate gland.

As with most tumors, prostate cancer develops through a multistage progression ultimately resulting in an aggressive tumor phenotype. The initial step in tumor progression involves the hyperproliferation of normal luminal and/or basal epithelial cells. Androgen responsive cells become hyperplastic and evolve into early-stage tumors. Although early-stage tumors are often androgen sensitive and respond to androgen ablation, a population of androgen independent cells evolve from the hyperplastic population. These cells represent a more advanced form of prostate tumor that may become invasive and potentially become metastatic to the bone, brain, or lung. A variety of genes may be differentially expressed during tumor progression. For example, loss of heterozygosity (LOH) is frequently observed on chromosome 8p in prostate cancer. Fluorescence *in situ* hybridization (FISH) revealed a deletion for at least 1 locus on 8p in 29 (69%) tumors, with a significantly higher frequency of the deletion on 8p21.2-p21.1 in advanced prostate cancer than in localized prostate cancer, implying that deletions on 8p22-p21.3 play an important role in tumor differentiation, while

8p21.2-p21.1 deletion plays a role in progression of prostate cancer (Oba, K. et al. (2001) Cancer Genet. Cytogenet. 124: 20-26).

A primary diagnostic marker for prostate cancer is prostate specific antigen (PSA). PSA is a tissue-specific serine protease almost exclusively produced by prostatic epithelial cells. The quantity of PSA correlates with the number and volume of the prostatic epithelial cells, and consequently, the levels of PSA are an excellent indicator of abnormal prostate growth. Men with prostate cancer exhibit an early linear increase in PSA levels followed by an exponential increase prior to diagnosis. However, since PSA levels are also influenced by factors such as inflammation, androgen and other growth factors, some scientists maintain that changes in PSA levels are not useful in detecting individual cases of prostate cancer.

Current areas of cancer research provide additional prospects for markers as well as potential therapeutic targets for prostate cancer. Several growth factors have been shown to play a critical role in tumor development, growth, and progression. The growth factors Epidermal Growth Factor (EGF), Fibroblast Growth Factor (FGF), and Tumor Growth Factor alpha (TGFα) are important in the growth of normal as well as hyperproliferative prostate epithelial cells, particularly at early stages of tumor development and progression, and affect signaling pathways in these cells in various ways (Lin, J. et al. (1999) Cancer Res. 59:2891-2897; Putz, T. et al. (1999) Cancer Res. 59:227-233). The TGF-β family of growth factors are generally expressed at increased levels in human cancers and the high expression levels in many cases correlates with advanced stages of malignancy and poor survival (Gold, L.I. (1999) Crit. Rev. Oncog. 10:303-360). Finally, there are human cell lines representing both the androgen-dependent stage of prostate cancer (LNCap) as well as the androgen-independent, hormone refractory stage of the disease (PC3 and DU-145) that have proved useful in studying gene expression patterns associated with the progression of prostate cancer, and the effects of cell treatments on these expressed genes (Chung, T.D. (1999) Prostate 15:199-207).

<u>Obesity</u>

The most important function of adipose tissue is its ability to store and release fat during periods of feeding and fasting. White adipose tissue is the major energy reserve in periods of excess energy use. Its primary purpose is mobilization during energy deprivation. Understanding how various molecules regulate adiposity and energy balance in physiological and pathophysiological situations may lead to the development of novel therapeutics for human obesity. Adipose tissue is also one of the important target tissues for insulin. Adipogenesis and insulin resistance in type II diabetes are linked and present intriguing relations. Most patients with type II diabetes are obese and obesity in turn

causes insulin resistance.

The majority of research in adipocyte biology to date has been done using transformed mouse preadipocyte cell lines. The culture condition which stimulates mouse preadipocyte differentiation is different from that for inducing human primary preadipocyte differentiation. In addition, primary cells are diploid and may therefore reflect the *in vivo* context better than aneuploid cell lines. Understanding the gene expression profile during adipogenesis in humans will lead to an understanding of the fundamental mechanism of adiposity regulation. Furthermore, through comparing the gene expression profiles of adipogenesis between donor with normal weight and donor with obesity, identification of crucial genes, potential drug targets for obesity and type II diabetes, will be possible.

Thiazolidinediones (TZDs) act as agonists for the peroxisome-proliferator-activated receptor gamma (PPARγ), a member of the nuclear hormone receptor superfamily. TZDs reduce hyperglycemia, hyperinsulinemia, and hypertension, in part by promoting glucose metabolism and inhibiting gluconeogenesis. Roles for PPARγ and its agonists have been demonstrated in a wide range of pathological conditions including diabetes, obesity, hypertension, atherosclerosis, polycystic ovarian syndrome, and cancers such as breast, prostate, liposarcoma, and colon cancer.

The mechanism by which TZDs and other PPARγ agonists enhance insulin sensitivity is not fully understood, but may involve the ability of PPARγ to promote adipogenesis. When ectopically expressed in cultured preadipocytes, PPARγ is a potent inducer of adipocyte differentiation. TZDs, in combination with insulin and other factors, can also enhance differentiation of human preadipocytes in culture (Adams et al. (1997) J. Clin. Invest. 100:3149-3153). The relative potency of different TZDs in promoting adipogenesis *in vitro* is proportional to both their insulin sensitizing effects *in vivo*, and their ability to bind and activate PPARγ *in vitro*. Interestingly, adipocytes derived from omental adipose depots are refractory to the effects of TZDs. It has therefore been suggested that the insulin sensitizing effects of TZDs may result from their ability to promote adipogenesis in subcutaneous adipose depots (Adams et al., *supra*). Further, dominant negative mutations in the PPARγ gene have been identified in two non-obese subjects with severe insulin resistance, hypertension, and overt non-insulin dependent diabetes mellitus (NIDDM) (Barroso et al. (1998) Nature 402:880-883).

NIDDM is the most common form of diabetes mellitus, a chronic metabolic disease that affects 143 million people worldwide. NIDDM is characterized by abnormal glucose and lipid metabolism that results from a combination of peripheral insulin resistance and defective insulin secretion. NIDDM has a complex, progressive etiology and a high degree of heritability. Numerous complications of diabetes including heart disease, stroke, renal failure, retinopathy, and peripheral neuropathy contribute to the high rate of morbidity and mortality.

At the molecular level, PPARγ functions as a ligand activated transcription factor. In the presence of ligand, PPARγ forms a heterodimer with the retinoid X receptor (RXR) which then activates transcription of target genes containing one or more copies of a PPARγ response element (PPRE). Many genes important in lipid storage and metabolism contain PPREs and have been identified as PPARγ targets, including PEPCK, aP2, LPL, ACS, and FAT-P (Auwerx, J. (1999) Diabetologia 42:1033-1049). Multiple ligands for PPARγ have been identified. These include a variety of fatty acid metabolites; synthetic drugs belonging to the TZD class, such as Pioglitazone and Rosiglitazone (BRL49653); and certain non-glitazone tyrosine analogs such as GI262570 and GW1929. The prostaglandin derivative 15-dPGJ2 is a potent endogenous ligand for PPARγ.

Expression of PPARγ is very high in adipose but barely detectable in skeletal muscle, the primary site for insulin stimulated glucose disposal in the body. PPARγ is also moderately expressed in large intestine, kidney, liver, vascular smooth muscle, hematopoietic cells, and macrophages. The high expression of PPARγ in adipose tissue suggests that the insulin sensitizing effects of TZDs may result from alterations in the expression of one or more PPARγ regulated genes in adipose tissue. Identification of PPARγ target genes will contribute to better drug design and the development of novel therapeutic strategies for diabetes, obesity, and other conditions.

Systematic attempts to identify PPARγ target genes have been made in several rodent models of obesity and diabetes (Suzuki et al. (2000) Jpn. J. Pharmacol. 84:113-123; Way et al. (2001) Endocrinology 142:1269-1277). However, a serious drawback of the rodent gene expression studies is that significant differences exist between human and rodent models of adipogenesis, diabetes, and obesity (Taylor (1999) Cell 97:9-12; Gregoire et al. (1998) Physiol. Reviews 78:783-809). Therefore, an unbiased approach to identifying TZD regulated genes in primary cultures of human tissues is necessary to fully elucidate the molecular basis for diseases associated with PPARγ activity.

Ovarian Cancer

Ovarian cancer is the leading cause of death from a gynecologic cancer. The majority of ovarian cancers are derived from epithelial cells, and 70% of patients with epithelial ovarian cancers present with late-stage disease. As a result, the long-term survival rate for this disease is very low. Identification of early-stage markers for ovarian cancer would significantly increase the survival rate. Genetic variations involved in ovarian cancer development include mutation of p53 and microsatellite instability. Gene expression patterns likely vary when normal ovary is compared to ovarian tumors.

Tangier Disease

Tangier disease (TD) is a genetic disorder characterized by the near absence of circulating high density lipoprotein (HDL) and the accumulation of cholesterol esters in many tissues, including

tonsils, lymph nodes, liver, spleen, thymus, and intestine. Low levels of HDL represent a clear predictor of premature coronary artery disease and homozygous TD correlates with a four- to six-fold increase in cardiovascular disease compared to controls. HDL plays a cardio-protective role in reverse cholesterol transport, the flux of cholesterol from peripheral cells such as tissue macrophages through plasma lipoproteins to the liver. The HDL protein, apolipoprotein A-I, plays a major role in this process, interacting with the cell surface to remove excess cholesterol and phospholipids. This pathway is severely impaired in TD and the defect lies in a specific gene, the ABC1 transporter. This gene is a member of the family of ATP-binding cassette transporters, which utilize ATP hydrolysis to transport a variety of substrates across membranes.

Human Endothelium

Human ECV304 cells are an immortalized endothelial cell line that grows without external stimulus. ECV304s have been used as an experimental model for investigating *in vitro* the role of endothelium in human vascular biology. Activation of vascular endothelium is considered to be a central event in a wide range of both physiological and pathophysiological processes, such as vascular tone regulation, coagulation and thrombosis, atherosclerosis, and inflammation.

Inflammatory Response

TNF-α is a pleiotropic cytokine that plays a central role in mediating the inflammatory response through activation of multiple signal transduction pathways. TNF-α is produced by activated lymphocytes, macrophages, and other white blood cells and can activate endothelial cells. Monitoring the endothelial cells' response to TNF-α at the level of mRNA expression can provide information necessary for better understanding of both TNF-α signaling pathways and endothelial cell biology.

Gemfibrozil

Gemfibrozil is a fibric acid antilipemic agent that lowers serum triglycerides and produces favorable changes in lipoproteins. Gemfibrozil is effective in reducing the risk of coronary heart disease in men (Frick, M.H., et al. (1987) New Engl. J. Med; 317:1237-1245). The compound can inhibit peripheral lipolysis and decrease hepatic extraction of free fatty acids, which decreases hepatic triglyceride production. Gemfibrozil also inhibits the synthesis and increases the clearance of apolipoprotein B, a carrier molecule for VLDL. Gemfibrozil has variable effects on LDL cholesterol. Although it causes moderate reductions in patients with type IIa hyperlipoproteinemia, changes in patients with either type IIb or type IV hyperlipoproteinemia are unpredictable. In general, the HMG-CoA reductase inhibitors are more effective than gemfibrozil in reducing LDL cholesterol. At the molecular level gemfibozil may function as a peroxisome proliferator-activated receptor (PPAR) agonist. Gemfibrozil is rapidly and completely absorbed from the GI tract and undergoes

enterohepatic recirculation. Gemfibrozil is metabolized by the liver and excreted by the kidneys, mainly as metabolites, one of which possesses pharmacologic activity. Gemfibozil causes peroxisome proliferation and hepatocarcinogenesis in rats, which is a cause for concern generally for fibric acid derivative drugs. In humans, fibric acid derivatives are known to increase the risk of gall bladder disease although gemfibrozil is better tolerated than other fibrates. The relative safety of gemfibrozil in humans compared to rodent species including rats may be attributed to differences in metabolism and clearance of the compound in different species (Dix, K.J., et al. (1999) Drug Metab. Distrib. 27:138-146; Thomas, B.F., et al. (1999) Drug Metab. Distrib. 27:147-157).

C3A Cell Line

The human C3A cell line is a clonal derivative of HepG2/C3 (hepatoma cell line, isolated from a 15-year-old male with liver tumor), which was selected for strong contact inhibition of growth. The use of a clonal population enhances the reproducibility of the cells. C3A cells have many characteristics of primary human hepatocytes in culture: i) expression of insulin receptor and insulin-like growth factor II receptor; ii) secretion of a high ratio of serum albumin compared with α-fetoprotein; iii) conversion of ammonia to urea and glutamine; iv) metabolism of aromatic amino acids; and v) proliferation in glucose-free and insulin-free medium. The C3A cell line is now well established as an *in vitro* model of the mature human liver (Mickelson et al. (1995) Hepatology 22:866-875; Nagendra et al. (1997) Am. J. Physiol. 272:G408-G416).

Lung Cancer

Lung cancer is the leading cause of cancer death in the United States, affecting more than 100,000 men and 50,000 women each year. Nearly 90% of the patients diagnosed with lung cancer are cigarette smokers. Tobacco smoke contains thousands of noxious substances that induce carcinogen metabolizing enzymes and covalent DNA adduct formation in the exposed bronchial epithelium. In nearly 80% of patients diagnosed with lung cancer, metastasis has already occurred. Most commonly lung cancers metastasize to pleura, brain, bone, pericardium, and liver. The decision to treat with surgery, radiation therapy, or chemotherapy is made on the basis of tumor histology, response to growth factors or hormones, and sensitivity to inhibitors or drugs. With current treatments, most patients die within one year of diagnosis. Earlier diagnosis and a systematic approach to identification, staging, and treatment of lung cancer could positively affect patient outcome.

Lung cancers progress through a series of morphologically distinct stages from hyperplasia to invasive carcinoma. Malignant lung cancers are divided into two groups comprising four histopathological classes. The Non Small Cell Lung Carcinoma (NSCLC) group includes squamous

cell carcinomas, adenocarcinomas, and large cell carcinomas and accounts for about 70% of all lung cancer cases. Adenocarcinomas typically arise in the peripheral airways and often form mucin secreting glands. Squamous cell carcinomas typically arise in proximal airways. The histogenesis of squamous cell carcinomas may be related to chronic inflammation and injury to the bronchial epithelium, leading to squamous metaplasia. The Small Cell Lung Carcinoma (SCLC) group accounts for about 20% of lung cancer cases. SCLCs typically arise in proximal airways and exhibit a number of paraneoplastic syndromes including inappropriate production of adrenocorticotropin and anti-diuretic hormone.

Lung cancer cells accumulate numerous genetic lesions, many of which are associated with cytologically visible chromosomal aberrations. The high frequency of chromosomal deletions associated with lung cancer may reflect the role of multiple tumor suppressor loci in the etiology of this disease. Deletion of the short arm of chromosome 3 is found in over 90% of cases and represents one of the earliest genetic lesions leading to lung cancer. Deletions at chromosome arms 9p and 17p are also common. Other frequently observed genetic lesions include overexpression of telomerase, activation of oncogenes such as K-ras and c-myc, and inactivation of tumor suppressor genes such as RB, p53 and CDKN2.

Genes differentially regulated in lung cancer have been identified by a variety of methods. Using mRNA differential display technology, Manda et al. (1999; Genomics 51:5-14) identified five genes differentially expressed in lung cancer cell lines compared to normal bronchial epithelial cells. Among the known genes, pulmonary surfactant apoprotein A and alpha 2 macroglobulin were down regulated whereas nm23H1 was upregulated. Petersen et al.. (2000; Int J. Cancer, 86:512-517) used suppression subtractive hybridization to identify 552 clones differentially expressed in lung tumor derived cell lines, 205 of which represented known genes. Among the known genes, thrombospondin-1, fibronectin, intercellular adhesion molecule 1, and cytokeratins 6 and 18 were previously observed to be differentially expressed in lung cancers. Wang et al. (2000; Oncogene 19:1519-1528) used a combination of microarray analysis and subtractive hybridization to identify 17 genes differentially overexpressed in squamous cell carcinoma compared with normal lung epithelium. Among the known genes they identified were keratin isoform 6, KOC, SPRC, IGFb2, connexin 26, plakofillin 1 and cytokeratin 13.

T Cells

T cells require two distinct signals to achieve optimal activation. First, the "antigenic" signal delivered through the binding of the TCR-CD3 complex. Second, the costimulatory signal delivered through the binding of the CD28 molecules. Upon binding of the TCR-CD3 complex alone, T cells

only achieve a partial state of activation. However, it is important to note that the signaling requirements of T cell depend greatly on the cycling state of those cells.

PMA is a broad activator of the protein kinase C-dependent pathways. Ionomycin is a calcium ionophore that permits the entry of calcium in the cell, hence increasing the cytosolic calcium concentration. The combination of PMA and ionomycin activates two of the major signaling pathways used by mammalian cells to interact with their environment. In T cells, the combination of PMA and ionomycin mimics the type of secondary signaling events elicited during optimal B cell activation.

Colon Cancer

While soft tissue sarcomas are relatively rare, more than 50% of new patients diagnosed with the disease will die from it. The molecular pathways leading to the development of sarcomas are relatively unknown, due to the rarity of the disease and variation in pathology. Colon cancer evolves through a multi-step process whereby pre-malignant colonocytes undergo a relatively defined sequence of events leading to tumor formation. Several factors participate in the process of tumor progression and malignant transformation including genetic factors, mutations, and selection.

To understand the nature of gene alterations in colorectal cancer, a number of studies have focused on the inherited syndromes. Familial adenomatous polyposis (FAP), is caused by mutations in the adenomatous polyposis coli gene (APC), resulting in truncated or inactive forms of the protein. This tumor suppressor gene has been mapped to chromosome 5q. Hereditary nonpolyposis colorectal cancer (HNPCC) is caused by mutations in mis-match repair genes. Although hereditary colon cancer syndromes occur in a small percentage of the population and most colorectal cancers are considered sporadic, knowledge from studies of the hereditary syndromes can be generally applied. For instance, somatic mutations in APC occur in at least 80% of sporadic colon tumors. APC mutations are thought to be the initiating event in the disease. Other mutations occur subsequently. Approximately 50% of colorectal cancers contain activating mutations in ras, while 85% contain inactivating mutations in p53. Changes in all of these genes lead to gene expression changes in colon cancer.

There is a need in the art for new compositions, including nucleic acids and proteins, for the diagnosis, prevention, and treatment of immune system disorders, neurological disorders, developmental disorders, connective tissue disorders, and cell proliferative disorders, including cancer.

SUMMARY OF THE INVENTION

Various embodiments of the invention provide purified polypeptides, cell adhesion and extracellular matrix proteins, referred to collectively as 'CADECM' and individually as 'CADECM-1,'

'CADECM-2,' 'CADECM-3,' 'CADECM-4,' 'CADECM-5,' 'CADECM-6,' 'CADECM-7,' 'CADECM-8,' 'CADECM-9,' 'CADECM-10,' 'CADECM-11,' 'CADECM-12,' 'CADECM-13,' 'CADECM-14,' 'CADECM-15,' 'CADECM-16,' 'CADECM-17,' 'CADECM-18,' 'CADECM-19,' 'CADECM-20,' 'CADECM-21,' 'CADECM-22,' 'CADECM-23,' 'CADECM-24,' 'CADECM-25,' 'CADECM-26,' 'CADECM-27,' 'CADECM-28,' 'CADECM-29,' 'CADECM-30,' 'CADECM-31,' 'CADECM-32,' 'CADECM-33,' 'CADECM-34,' 'CADECM-35,' 'CADECM-36,' 'CADECM-37,' 'CADECM-38,' 'CADECM-39,' 'CADECM-40,' 'CADECM-41,' and 'CADECM-42' and methods for using these proteins and their encoding polynucleotides for the detection, diagnosis, and treatment of diseases and medical conditions. Embodiments also provide methods for utilizing the purified cell adhesion and extracellular matrix proteins and/or their encoding polynucleotides for facilitating the drug discovery process, including determination of efficacy, dosage, toxicity, and pharmacology. Related embodiments provide methods for utilizing the purified cell adhesion and extracellular matrix proteins and/or their encoding polynucleotides for investigating the pathogenesis of diseases and medical conditions.

An embodiment provides an isolated polypeptide selected from the group consisting of a) a polypeptide comprising an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, b) a polypeptide comprising a naturally occurring amino acid sequence at least 90% identical or at least about 90% identical to an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, c) a biologically active fragment of a polypeptide having an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, and d) an immunogenic fragment of a polypeptide having an amino acid sequence selected from the group consisting of SEQ ID NO:1-42. Another embodiment provides an isolated polypeptide comprising an amino acid sequence of SEQ ID NO:1-42.

Still another embodiment provides an isolated polynucleotide encoding a polypeptide selected from the group consisting of a) a polypeptide comprising an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, b) a polypeptide comprising a naturally occurring amino acid sequence at least 90% identical or at least about 90% identical to an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, c) a biologically active fragment of a polypeptide having an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, and d) an immunogenic fragment of a polypeptide having an amino acid sequence selected from the group consisting of SEQ ID NO:1-42. In another embodiment, the polynucleotide encodes a polypeptide selected from the group consisting of SEQ ID NO:1-42. In an alternative embodiment, the polynucleotide is selected from the group consisting of SEQ ID NO:43-84.

Still another embodiment provides a recombinant polynucleotide comprising a promoter

sequence operably linked to a polynucleotide encoding a polypeptide selected from the group consisting of a) a polypeptide comprising an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, b) a polypeptide comprising a naturally occurring amino acid sequence at least 90% identical or at least about 90% identical to an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, c) a biologically active fragment of a polypeptide having an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, and d) an immunogenic fragment of a polypeptide having an amino acid sequence selected from the group consisting of SEQ ID NO:1-42. Another embodiment provides a cell transformed with the recombinant polynucleotide. Yet another embodiment provides a transgenic organism comprising the recombinant polynucleotide.

Another embodiment provides a method for producing a polypeptide selected from the group consisting of a) a polypeptide comprising an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, b) a polypeptide comprising a naturally occurring amino acid sequence at least 90% identical or at least about 90% identical to an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, c) a biologically active fragment of a polypeptide having an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, and d) an immunogenic fragment of a polypeptide having an amino acid sequence selected from the group consisting of SEQ ID NO:1-42. The method comprises a) culturing a cell under conditions suitable for expression of the polypeptide, wherein said cell is transformed with a recombinant polynucleotide comprising a promoter sequence operably linked to a polynucleotide encoding the polypeptide, and b) recovering the polypeptide so expressed.

Yet another embodiment provides an isolated antibody which specifically binds to a polypeptide selected from the group consisting of a) a polypeptide comprising an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, b) a polypeptide comprising a naturally occurring amino acid sequence at least 90% identical or at least about 90% identical to an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, c) a biologically active fragment of a polypeptide having an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, and d) an immunogenic fragment of a polypeptide having an amino acid sequence selected from the group consisting of SEQ ID NO:1-42.

Still yet another embodiment provides an isolated polynucleotide selected from the group consisting of a) a polynucleotide comprising a polynucleotide sequence selected from the group consisting of SEQ ID NO:43-84, b) a polynucleotide comprising a naturally occurring polynucleotide sequence at least 90% identical or at least about 90% identical to a polynucleotide sequence selected from the group consisting of SEQ ID NO:43-84, c) a polynucleotide complementary to the

polynucleotide of a), d) a polynucleotide complementary to the polynucleotide of b), and e) an RNA equivalent of a)-d). In other embodiments, the polynucleotide can comprise at least about 20, 30, 40, 60, 80, or 100 contiguous nucleotides.

Yet another embodiment provides a method for detecting a target polynucleotide in a sample, said target polynucleotide being selected from the group consisting of a) a polynucleotide comprising a polynucleotide sequence selected from the group consisting of SEQ ID NO:43-84, b) a polynucleotide comprising a naturally occurring polynucleotide sequence at least 90% identical or at least about 90% identical to a polynucleotide sequence selected from the group consisting of SEQ ID NO:43-84, c) a polynucleotide complementary to the polynucleotide of a), d) a polynucleotide complementary to the polynucleotide of b), and e) an RNA equivalent of a)-d). The method comprises a) hybridizing the sample with a probe comprising at least 20 contiguous nucleotides comprising a sequence complementary to said target polynucleotide in the sample, and which probe specifically hybridizes to said target polynucleotide, under conditions whereby a hybridization complex is formed between said probe and said target polynucleotide or fragments thereof, and b) detecting the presence or absence of said hybridization complex. In a related embodiment, the method can include detecting the amount of the hybridization complex. In still other embodiments, the probe can comprise at least about 20, 30, 40, 60, 80, or 100 contiguous nucleotides.

Still yet another embodiment provides a method for detecting a target polynucleotide in a sample, said target polynucleotide being selected from the group consisting of a) a polynucleotide comprising a polynucleotide sequence selected from the group consisting of SEQ ID NO:43-84, b) a polynucleotide comprising a naturally occurring polynucleotide sequence at least 90% identical or at least about 90% identical to a polynucleotide sequence selected from the group consisting of SEQ ID NO:43-84, c) a polynucleotide complementary to the polynucleotide of a), d) a polynucleotide complementary to the polynucleotide of b), and e) an RNA equivalent of a)-d). The method comprises a) amplifying said target polynucleotide or fragment thereof using polymerase chain reaction amplification, and b) detecting the presence or absence of said amplified target polynucleotide or fragment thereof. In a related embodiment, the method can include detecting the amount of the amplified target polynucleotide or fragment thereof.

Another embodiment provides a composition comprising an effective amount of a polypeptide selected from the group consisting of a) a polypeptide comprising an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, b) a polypeptide comprising a naturally occurring amino acid sequence at least 90% identical or at least about 90% identical to an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, c) a biologically active fragment of a

polypeptide having an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, and d) an immunogenic fragment of a polypeptide having an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, and a pharmaceutically acceptable excipient. In one embodiment, the composition can comprise an amino acid sequence selected from the group consisting of SEQ ID NO:1-42. Other embodiments provide a method of treating a disease or condition associated with decreased or abnormal expression of functional CADECM, comprising administering to a patient in need of such treatment the composition.

Yet another embodiment provides a method for screening a compound for effectiveness as an agonist of a polypeptide selected from the group consisting of a) a polypeptide comprising an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, b) a polypeptide comprising a naturally occurring amino acid sequence at least 90% identical or at least about 90% identical to an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, c) a biologically active fragment of a polypeptide having an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, and d) an immunogenic fragment of a polypeptide having an amino acid sequence selected from the group consisting of SEQ ID NO:1-42. The method comprises a) contacting a sample comprising the polypeptide with a compound, and b) detecting agonist activity in the sample. Another embodiment provides a composition comprising an agonist compound identified by the method and a pharmaceutically acceptable excipient. Yet another embodiment provides a method of treating a disease or condition associated with decreased expression of functional CADECM, comprising administering to a patient in need of such treatment the composition.

Still yet another embodiment provides a method for screening a compound for effectiveness as an antagonist of a polypeptide selected from the group consisting of a) a polypeptide comprising an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, b) a polypeptide comprising a naturally occurring amino acid sequence at least 90% identical or at least about 90% identical to an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, c) a biologically active fragment of a polypeptide having an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, and d) an immunogenic fragment of a polypeptide having an amino acid sequence selected from the group consisting of SEQ ID NO:1-42. The method comprises a) contacting a sample comprising the polypeptide with a compound, and b) detecting antagonist activity in the sample. Another embodiment provides a composition comprising an antagonist compound identified by the method and a pharmaceutically acceptable excipient. Yet another embodiment provides a method of treating a disease or condition associated with overexpression of functional CADECM, comprising administering to a patient in need of such treatment the composition.

Another embodiment provides a method of screening for a compound that specifically binds to a polypeptide selected from the group consisting of a) a polypeptide comprising an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, b) a polypeptide comprising a naturally occurring amino acid sequence at least 90% identical or at least about 90% identical to an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, c) a biologically active fragment of a polypeptide having an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, and d) an immunogenic fragment of a polypeptide having an amino acid sequence selected from the group consisting of SEQ ID NO:1-42. The method comprises a) combining the polypeptide with at least one test compound under suitable conditions, and b) detecting binding of the polypeptide to the test compound, thereby identifying a compound that specifically binds to the polypeptide.

Yet another embodiment provides a method of screening for a compound that modulates the activity of a polypeptide selected from the group consisting of a) a polypeptide comprising an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, b) a polypeptide comprising a naturally occurring amino acid sequence at least 90% identical or at least about 90% identical to an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, c) a biologically active fragment of a polypeptide having an amino acid sequence selected from the group consisting of SEQ ID NO:1-42, and d) an immunogenic fragment of a polypeptide having an amino acid sequence selected from the group consisting of SEQ ID NO:1-42. The method comprises a) combining the polypeptide with at least one test compound under conditions permissive for the activity of the polypeptide, b) assessing the activity of the polypeptide in the presence of the test compound, and c) comparing the activity of the polypeptide in the presence of the test compound with the activity of the polypeptide in the absence of the test compound, wherein a change in the activity of the polypeptide in the presence of the test compound is indicative of a compound that modulates the activity of the polypeptide.

Still yet another embodiment provides a method for screening a compound for effectiveness in altering expression of a target polynucleotide, wherein said target polynucleotide comprises a polynucleotide sequence selected from the group consisting of SEQ ID NO:43-84, the method comprising a) contacting a sample comprising the target polynucleotide with a compound, b) detecting altered expression of the target polynucleotide, and c) comparing the expression of the target polynucleotide in the presence of varying amounts of the compound and in the absence of the compound.

Another embodiment provides a method for assessing toxicity of a test compound, said

method comprising a) treating a biological sample containing nucleic acids with the test compound; b) hybridizing the nucleic acids of the treated biological sample with a probe comprising at least 20 contiguous nucleotides of a polynucleotide selected from the group consisting of i) a polynucleotide comprising a polynucleotide sequence selected from the group consisting of SEQ ID NO:43-84, ii) a polynucleotide comprising a naturally occurring polynucleotide sequence at least 90% identical or at least about 90% identical to a polynucleotide sequence selected from the group consisting of SEQ ID NO:43-84, iii) a polynucleotide having a sequence complementary to i), iv) a polynucleotide complementary to the polynucleotide of ii), and v) an RNA equivalent of i)-iv). Hybridization occurs under conditions whereby a specific hybridization complex is formed between said probe and a target polynucleotide in the biological sample, said target polynucleotide selected from the group consisting of i) a polynucleotide comprising a polynucleotide sequence selected from the group consisting of SEQ ID NO:43-84, ii) a polynucleotide comprising a naturally occurring polynucleotide sequence at least 90% identical or at least about 90% identical to a polynucleotide sequence selected from the group consisting of SEQ ID NO:43-84, iii) a polynucleotide complementary to the polynucleotide of i), iv) a polynucleotide complementary to the polynucleotide of ii), and v) an RNA equivalent of i)-iv). Alternatively, the target polynucleotide can comprise a fragment of a polynucleotide selected from the group consisting of i)-v) above; c) quantifying the amount of hybridization complex; and d) comparing the amount of hybridization complex in the treated biological sample with the amount of hybridization complex in an untreated biological sample, wherein a difference in the amount of hybridization complex in the treated biological sample is indicative of toxicity of the test compound.

BRIEF DESCRIPTION OF THE TABLES

Table 1 summarizes the nomenclature for full length polynucleotide and polypeptide embodiments of the invention.

Table 2 shows the GenBank identification number and annotation of the nearest GenBank homolog, and the PROTEOME database identification numbers and annotations of PROTEOME database homologs, for polypeptide embodiments of the invention. The probability scores for the matches between each polypeptide and its homolog(s) are also shown.

Table 3 shows structural features of polypeptide embodiments, including predicted motifs and domains, along with the methods, algorithms, and searchable databases used for analysis of the polypeptides.

Table 4 lists the cDNA and/or genomic DNA fragments which were used to assemble polynucleotide embodiments, along with selected fragments of the polynucleotides.

Table 5 shows representative cDNA libraries for polynucleotide embodiments.

Table 6 provides an appendix which describes the tissues and vectors used for construction of the cDNA libraries shown in Table 5.

Table 7 shows the tools, programs, and algorithms used to analyze polynucleotides and polypeptides, along with applicable descriptions, references, and threshold parameters.

Table 8 shows single nucleotide polymorphisms found in polynucleotide sequences of the invention, along with allele frequencies in different human populations.

DESCRIPTION OF THE INVENTION

Before the present proteins, nucleic acids, and methods are described, it is understood that embodiments of the invention are not limited to the particular machines, instruments, materials, and methods described, as these may vary. It is also to be understood that the terminology used herein is for the purpose of describing particular embodiments only, and is not intended to limit the scope of the invention.

As used herein and in the appended claims, the singular forms "a," "an," and "the" include plural reference unless the context clearly dictates otherwise. Thus, for example, a reference to "a host cell" includes a plurality of such host cells, and a reference to "an antibody" is a reference to one or more antibodies and equivalents thereof known to those skilled in the art, and so forth.

Unless defined otherwise, all technical and scientific terms used herein have the same meanings as commonly understood by one of ordinary skill in the art to which this invention belongs. Although any machines, materials, and methods similar or equivalent to those described herein can be used to practice or test the present invention, the preferred machines, materials and methods are now described. All publications mentioned herein are cited for the purpose of describing and disclosing the cell lines, protocols, reagents and vectors which are reported in the publications and which might be used in connection with various embodiments of the invention. Nothing herein is to be construed as an admission that the invention is not entitled to antedate such disclosure by virtue of prior invention.

DEFINITIONS

"CADECM" refers to the amino acid sequences of substantially purified CADECM obtained from any species, particularly a mammalian species, including bovine, ovine, porcine, murine, equine, and human, and from any source, whether natural, synthetic, semi-synthetic, or recombinant.

The term "agonist" refers to a molecule which intensifies or mimics the biological activity of CADECM. Agonists may include proteins, nucleic acids, carbohydrates, small molecules, or any other compound or composition which modulates the activity of CADECM either by directly interacting with

CADECM or by acting on components of the biological pathway in which CADECM participates.

An "allelic variant" is an alternative form of the gene encoding CADECM. Allelic variants may result from at least one mutation in the nucleic acid sequence and may result in altered mRNAs or in polypeptides whose structure or function may or may not be altered. A gene may have none, one, or many allelic variants of its naturally occurring form. Common mutational changes which give rise to allelic variants are generally ascribed to natural deletions, additions, or substitutions of nucleotides. Each of these types of changes may occur alone, or in combination with the others, one or more times in a given sequence.

"Altered" nucleic acid sequences encoding CADECM include those sequences with deletions, insertions, or substitutions of different nucleotides, resulting in a polypeptide the same as CADECM or a polypeptide with at least one functional characteristic of CADECM. Included within this definition are polymorphisms which may or may not be readily detectable using a particular oligonucleotide probe of the polynucleotide encoding CADECM, and improper or unexpected hybridization to allelic variants, with a locus other than the normal chromosomal locus for the polynucleotide encoding CADECM. The encoded protein may also be "altered," and may contain deletions, insertions, or substitutions of amino acid residues which produce a silent change and result in a functionally equivalent CADECM. Deliberate amino acid substitutions may be made on the basis of one or more similarities in polarity, charge, solubility, hydrophobicity, hydrophilicity, and/or the amphipathic nature the residues, as long as the biological or immunological activity of CADECM is retained. For example negatively charged amino acids may include aspartic acid and glutamic acid, and positively charged amino acids may include lysine and arginine. Amino acids with uncharged polar side chains having similar hydrophilicity values may include: asparagine and glutamine; and serine and threonine. Amino acids with uncharged side chains having similar hydrophilicity values may include: leucine, isoleucine, and valine; glycine and alanine; and phenylalanine and tyrosine.

The terms "amino acid" and "amino acid sequence" can refer to an oligopeptide, a peptide, a polypeptide, or a protein sequence, or a fragment of any of these, and to naturally occurring or synthetic molecules. Where "amino acid sequence" is recited to refer to a sequence of a naturally occurring protein molecule, "amino acid sequence" and like terms are not meant to limit the amino acid sequence to the complete native amino acid sequence associated with the recited protein molecule.

"Amplification" relates to the production of additional copies of a nucleic acid. Amplification may be carried out using polymerase chain reaction (PCR) technologies or other nucleic acid amplification technologies well known in the art.

The term "antagonist" refers to a molecule which inhibits or attenuates the biological activity

of CADECM. Antagonists may include proteins such as antibodies, anticalins, nucleic acids, carbohydrates, small molecules, or any other compound or composition which modulates the activity of CADECM either by directly interacting with CADECM or by acting on components of the biological pathway in which CADECM participates.

The term "antibody" refers to intact immunoglobulin molecules as well as to fragments thereof, such as Fab, $F(ab')_2$, and Fv fragments, which are capable of binding an epitopic determinant. Antibodies that bind CADECM polypeptides can be prepared using intact polypeptides or using fragments containing small peptides of interest as the immunizing antigen. The polypeptide or oligopeptide used to immunize an animal (e.g., a mouse, a rat, or a rabbit) can be derived from the translation of RNA, or synthesized chemically, and can be conjugated to a carrier protein if desired. Commonly used carriers that are chemically coupled to peptides include bovine serum albumin, thyroglobulin, and keyhole limpet hemocyanin (KLH). The coupled peptide is then used to immunize the animal.

The term "antigenic determinant" refers to that region of a molecule (i.e., an epitope) that makes contact with a particular antibody. When a protein or a fragment of a protein is used to immunize a host animal, numerous regions of the protein may induce the production of antibodies which bind specifically to antigenic determinants (particular regions or three-dimensional structures on the protein). An antigenic determinant may compete with the intact antigen (i.e., the immunogen used to elicit the immune response) for binding to an antibody.

The term "aptamer" refers to a nucleic acid or oligonucleotide molecule that binds to a specific molecular target. Aptamers are derived from an *in vitro* evolutionary process (e.g., SELEX (Systematic Evolution of Ligands by EXponential Enrichment), described in U.S. Patent No. 5,270,163), which selects for target-specific aptamer sequences from large combinatorial libraries. Aptamer compositions may be double-stranded or single-stranded, and may include deoxyribonucleotides, ribonucleotides, nucleotide derivatives, or other nucleotide-like molecules. The nucleotide components of an aptamer may have modified sugar groups (e.g., the 2'-OH group of a ribonucleotide may be replaced by 2'-F or $2'\text{-}NH_2$), which may improve a desired property, e.g., resistance to nucleases or longer lifetime in blood. Aptamers may be conjugated to other molecules, e.g., a high molecular weight carrier to slow clearance of the aptamer from the circulatory system. Aptamers may be specifically cross-linked to their cognate ligands, e.g., by photo-activation of a cross-linker (Brody, E.N. and L. Gold (2000) J. Biotechnol. 74:5-13).

The term "intramer" refers to an aptamer which is expressed *in vivo*. For example, a vaccinia virus-based RNA expression system has been used to express specific RNA aptamers at

high levels in the cytoplasm of leukocytes (Blind, M. et al. (1999) Proc. Natl. Acad. Sci. USA 96:3606-3610).

The term "spiegelmer" refers to an aptamer which includes L-DNA, L-RNA, or other left-handed nucleotide derivatives or nucleotide-like molecules. Aptamers containing left-handed nucleotides are resistant to degradation by naturally occurring enzymes, which normally act on substrates containing right-handed nucleotides.

The term "antisense" refers to any composition capable of base-pairing with the "sense" (coding) strand of a polynucleotide having a specific nucleic acid sequence. Antisense compositions may include DNA; RNA; peptide nucleic acid (PNA); oligonucleotides having modified backbone linkages such as phosphorothioates, methylphosphonates, or benzylphosphonates; oligonucleotides having modified sugar groups such as 2'-methoxyethyl sugars or 2'-methoxyethoxy sugars; or oligonucleotides having modified bases such as 5-methyl cytosine, 2'-deoxyuracil, or 7-deaza-2'-deoxyguanosine. Antisense molecules may be produced by any method including chemical synthesis or transcription. Once introduced into a cell, the complementary antisense molecule base-pairs with a naturally occurring nucleic acid sequence produced by the cell to form duplexes which block either transcription or translation. The designation "negative" or "minus" can refer to the antisense strand, and the designation "positive" or "plus" can refer to the sense strand of a reference DNA molecule.

The term "biologically active" refers to a protein having structural, regulatory, or biochemical functions of a naturally occurring molecule. Likewise, "immunologically active" or "immunogenic" refers to the capability of the natural, recombinant, or synthetic CADECM, or of any oligopeptide thereof, to induce a specific immune response in appropriate animals or cells and to bind with specific antibodies.

"Complementary" describes the relationship between two single-stranded nucleic acid sequences that anneal by base-pairing. For example, 5'-AGT-3' pairs with its complement, 3'-TCA-5'.

A "composition comprising a given polynucleotide" and a "composition comprising a given polypeptide" can refer to any composition containing the given polynucleotide or polypeptide. The composition may comprise a dry formulation or an aqueous solution. Compositions comprising polynucleotides encoding CADECM or fragments of CADECM may be employed as hybridization probes. The probes may be stored in freeze-dried form and may be associated with a stabilizing agent such as a carbohydrate. In hybridizations, the probe may be deployed in an aqueous solution containing salts (e.g., NaCl), detergents (e.g., sodium dodecyl sulfate; SDS), and other components (e.g., Denhardt's solution, dry milk, salmon sperm DNA, etc.).

"Consensus sequence" refers to a nucleic acid sequence which has been subjected to repeated DNA sequence analysis to resolve uncalled bases, extended using the XL-PCR kit (Applied Biosystems, Foster City CA) in the 5' and/or the 3' direction, and resequenced, or which has been assembled from one or more overlapping cDNA, EST, or genomic DNA fragments using a computer program for fragment assembly, such as the GELVIEW fragment assembly system (Accelrys, Burlington MA) or Phrap (University of Washington, Seattle WA). Some sequences have been both extended and assembled to produce the consensus sequence.

"Conservative amino acid substitutions" are those substitutions that are predicted to least interfere with the properties of the original protein, i.e., the structure and especially the function of the protein is conserved and not significantly changed by such substitutions. The table below shows amino acids which may be substituted for an original amino acid in a protein and which are regarded as conservative amino acid substitutions.

Original Residue	Conservative Substitution
Ala	Gly, Ser
Arg	His, Lys
Asn	Asp, Gln, His
Asp	Asn, Glu
Cys	Ala, Ser
Gln	Asn, Glu, His
Glu	Asp, Gln, His
Gly	Ala
His	Asn, Arg, Gln, Glu
Ile	Leu, Val
Leu	Ile, Val
Lys	Arg, Gln, Glu
Met	Leu, Ile
Phe	His, Met, Leu, Trp, Tyr
Ser	Cys, Thr
Thr	Ser, Val
Trp	Phe, Tyr
Tyr	His, Phe, Trp
Val	Ile, Leu, Thr

Conservative amino acid substitutions generally maintain (a) the structure of the polypeptide backbone in the area of the substitution, for example, as a beta sheet or alpha helical conformation, (b) the charge or hydrophobicity of the molecule at the site of the substitution, and/or (c) the bulk of the side chain.

A "deletion" refers to a change in the amino acid or nucleotide sequence that results in the absence of one or more amino acid residues or nucleotides.

The term "derivative" refers to a chemically modified polynucleotide or polypeptide.

Chemical modifications of a polynucleotide can include, for example, replacement of hydrogen by an alkyl, acyl, hydroxyl, or amino group. A derivative polynucleotide encodes a polypeptide which retains at least one biological or immunological function of the natural molecule. A derivative polypeptide is one modified by glycosylation, pegylation, or any similar process that retains at least one biological or immunological function of the polypeptide from which it was derived.

A "detectable label" refers to a reporter molecule or enzyme that is capable of generating a measurable signal and is covalently or noncovalently joined to a polynucleotide or polypeptide.

"Differential expression" refers to increased or upregulated; or decreased, downregulated, or absent gene or protein expression, determined by comparing at least two different samples. Such comparisons may be carried out between, for example, a treated and an untreated sample, or a diseased and a normal sample.

"Exon shuffling" refers to the recombination of different coding regions (exons). Since an exon may represent a structural or functional domain of the encoded protein, new proteins may be assembled through the novel reassortment of stable substructures, thus allowing acceleration of the evolution of new protein functions.

A "fragment" is a unique portion of CADECM or a polynucleotide encoding CADECM which can be identical in sequence to, but shorter in length than, the parent sequence. A fragment may comprise up to the entire length of the defined sequence, minus one nucleotide/amino acid residue. For example, a fragment may comprise from about 5 to about 1000 contiguous nucleotides or amino acid residues. A fragment used as a probe, primer, antigen, therapeutic molecule, or for other purposes, may be at least 5, 10, 15, 16, 20, 25, 30, 40, 50, 60, 75, 100, 150, 250 or at least 500 contiguous nucleotides or amino acid residues in length. Fragments may be preferentially selected from certain regions of a molecule. For example, a polypeptide fragment may comprise a certain length of contiguous amino acids selected from the first 250 or 500 amino acids (or first 25% or 50%) of a polypeptide as shown in a certain defined sequence. Clearly these lengths are exemplary, and any length that is supported by the specification, including the Sequence Listing, tables, and figures, may be encompassed by the present embodiments.

A fragment of SEQ ID NO:43-84 can comprise a region of unique polynucleotide sequence that specifically identifies SEQ ID NO:43-84, for example, as distinct from any other sequence in the genome from which the fragment was obtained. A fragment of SEQ ID NO:43-84 can be employed in one or more embodiments of methods of the invention, for example, in hybridization and amplification technologies and in analogous methods that distinguish SEQ ID NO:43-84 from related polynucleotides. The precise length of a fragment of SEQ ID NO:43-84 and the region of SEQ ID

NO:43-84 to which the fragment corresponds are routinely determinable by one of ordinary skill in the art based on the intended purpose for the fragment.

A fragment of SEQ ID NO:1-42 is encoded by a fragment of SEQ ID NO:43-84. A fragment of SEQ ID NO:1-42 can comprise a region of unique amino acid sequence that specifically identifies SEQ ID NO:1-42. For example, a fragment of SEQ ID NO:1-42 can be used as an immunogenic peptide for the development of antibodies that specifically recognize SEQ ID NO:1-42. The precise length of a fragment of SEQ ID NO:1-42 and the region of SEQ ID NO:1-42 to which the fragment corresponds can be determined based on the intended purpose for the fragment using one or more analytical methods described herein or otherwise known in the art.

A "full length" polynucleotide is one containing at least a translation initiation codon (e.g., methionine) followed by an open reading frame and a translation termination codon. A "full length" polynucleotide sequence encodes a "full length" polypeptide sequence.

"Homology" refers to sequence similarity or, alternatively, sequence identity, between two or more polynucleotide sequences or two or more polypeptide sequences.

The terms "percent identity" and "% identity," as applied to polynucleotide sequences, refer to the percentage of identical nucleotide matches between at least two polynucleotide sequences aligned using a standardized algorithm. Such an algorithm may insert, in a standardized and reproducible way, gaps in the sequences being compared in order to optimize alignment between two sequences, and therefore achieve a more meaningful comparison of the two sequences.

Percent identity between polynucleotide sequences may be determined using one or more computer algorithms or programs known in the art or described herein. For example, percent identity can be determined using the default parameters of the CLUSTAL V algorithm as incorporated into the MEGALIGN version 3.12e sequence alignment program. This program is part of the LASERGENE software package, a suite of molecular biological analysis programs (DNASTAR, Madison WI). CLUSTAL V is described in Higgins, D.G. and P.M. Sharp (1989; CABIOS 5:151-153) and in Higgins, D.G. et al. (1992; CABIOS 8:189-191). For pairwise alignments of polynucleotide sequences, the default parameters are set as follows: Ktuple=2, gap penalty=5, window=4, and "diagonals saved"=4. The "weighted" residue weight table is selected as the default.

Alternatively, a suite of commonly used and freely available sequence comparison algorithms which can be used is provided by the National Center for Biotechnology Information (NCBI) Basic Local Alignment Search Tool (BLAST) (Altschul, S.F. et al. (1990) J. Mol. Biol. 215:403-410), which is available from several sources, including the NCBI, Bethesda, MD, and on the Internet at ncbi.nlm.nih.gov/BLAST/. The BLAST software suite includes various sequence analysis programs

including "blastn," that is used to align a known polynucleotide sequence with other polynucleotide sequences from a variety of databases. Also available is a tool called "BLAST 2 Sequences" that is used for direct pairwise comparison of two nucleotide sequences. "BLAST 2 Sequences" can be accessed and used interactively at ncbi.nlm.nih.gov/gorf/bl2.html. The "BLAST 2 Sequences" tool can be used for both blastn and blastp (discussed below). BLAST programs are commonly used with gap and other parameters set to default settings. For example, to compare two nucleotide sequences, one may use blastn with the "BLAST 2 Sequences" tool Version 2.0.12 (April-21-2000) set at default parameters. Such default parameters may be, for example:

Matrix: BLOSUM62

Reward for match: 1

Penalty for mismatch: -2

Open Gap: 5 and Extension Gap: 2 penalties

Gap x drop-off: 50

Expect: 10

Word Size: 11

Filter: on

Percent identity may be measured over the length of an entire defined sequence, for example, as defined by a particular SEQ ID number, or may be measured over a shorter length, for example, over the length of a fragment taken from a larger, defined sequence, for instance, a fragment of at least 20, at least 30, at least 40, at least 50, at least 70, at least 100, or at least 200 contiguous nucleotides. Such lengths are exemplary only, and it is understood that any fragment length supported by the sequences shown herein, in the tables, figures, or Sequence Listing, may be used to describe a length over which percentage identity may be measured.

Nucleic acid sequences that do not show a high degree of identity may nevertheless encode similar amino acid sequences due to the degeneracy of the genetic code. It is understood that changes in a nucleic acid sequence can be made using this degeneracy to produce multiple nucleic acid sequences that all encode substantially the same protein.

The phrases "percent identity" and "% identity," as applied to polypeptide sequences, refer to the percentage of identical residue matches between at least two polypeptide sequences aligned using a standardized algorithm. Methods of polypeptide sequence alignment are well-known. Some alignment methods take into account conservative amino acid substitutions. Such conservative substitutions, explained in more detail above, generally preserve the charge and hydrophobicity at the site of substitution, thus preserving the structure (and therefore function) of the polypeptide. The

phrases "percent similarity" and "% similarity," as applied to polypeptide sequences, refer to the percentage of residue matches, including identical residue matches and conservative substitutions, between at least two polypeptide sequences aligned using a standardized algorithm. In contrast, conservative substitutions are not included in the calculation of percent identity between polypeptide sequences.

Percent identity between polypeptide sequences may be determined using the default parameters of the CLUSTAL V algorithm as incorporated into the MEGALIGN version 3.12e sequence alignment program (described and referenced above). For pairwise alignments of polypeptide sequences using CLUSTAL V, the default parameters are set as follows: Ktuple=1, gap penalty=3, window=5, and "diagonals saved"=5. The PAM250 matrix is selected as the default residue weight table.

Alternatively the NCBI BLAST software suite may be used. For example, for a pairwise comparison of two polypeptide sequences, one may use the "BLAST 2 Sequences" tool Version 2.0.12 (April-21-2000) with blastp set at default parameters. Such default parameters may be, for example:

Matrix: BLOSUM62

Open Gap: 11 and Extension Gap: 1 penalties

Gap x drop-off: 50

Expect: 10

Word Size: 3

Filter: on

Percent identity may be measured over the length of an entire defined polypeptide sequence, for example, as defined by a particular SEQ ID number, or may be measured over a shorter length, for example, over the length of a fragment taken from a larger, defined polypeptide sequence, for instance, a fragment of at least 15, at least 20, at least 30, at least 40, at least 50, at least 70 or at least 150 contiguous residues. Such lengths are exemplary only, and it is understood that any fragment length supported by the sequences shown herein, in the tables, figures or Sequence Listing, may be used to describe a length over which percentage identity may be measured.

"Human artificial chromosomes" (HACs) are linear microchromosomes which may contain DNA sequences of about 6 kb to 10 Mb in size and which contain all of the elements required for chromosome replication, segregation and maintenance.

The term "humanized antibody" refers to an antibody molecule in which the amino acid sequence in the non-antigen binding regions has been altered so that the antibody more closely

resembles a human antibody, and still retains its original binding ability.

"Hybridization" refers to the process by which a polynucleotide strand anneals with a complementary strand through base pairing under defined hybridization conditions. Specific hybridization is an indication that two nucleic acid sequences share a high degree of complementarity. Specific hybridization complexes form under permissive annealing conditions and remain hybridized after the "washing" step(s). The washing step(s) is particularly important in determining the stringency of the hybridization process, with more stringent conditions allowing less non-specific binding, i.e., binding between pairs of nucleic acid strands that are not perfectly matched. Permissive conditions for annealing of nucleic acid sequences are routinely determinable by one of ordinary skill in the art and may be consistent among hybridization experiments, whereas wash conditions may be varied among experiments to achieve the desired stringency, and therefore hybridization specificity. Permissive annealing conditions occur, for example, at 68°C in the presence of about 6 x SSC, about 1% (w/v) SDS, and about 100 μg/ml sheared, denatured salmon sperm DNA.

Generally, stringency of hybridization is expressed, in part, with reference to the temperature under which the wash step is carried out. Such wash temperatures are typically selected to be about 5°C to 20°C lower than the thermal melting point (T_m) for the specific sequence at a defined ionic strength and pH. The T_m is the temperature (under defined ionic strength and pH) at which 50% of the target sequence hybridizes to a perfectly matched probe. An equation for calculating T_m and conditions for nucleic acid hybridization are well known and can be found in Sambrook, J. and D.W. Russell (2001; <u>Molecular Cloning: A Laboratory Manual</u>, 3rd ed., vol. 1-3, Cold Spring Harbor Press, Cold Spring Harbor NY, ch. 9).

High stringency conditions for hybridization between polynucleotides of the present invention include wash conditions of 68°C in the presence of about 0.2 x SSC and about 0.1% SDS, for 1 hour. Alternatively, temperatures of about 65°C, 60°C, 55°C, or 42°C may be used. SSC concentration may be varied from about 0.1 to 2 x SSC, with SDS being present at about 0.1%. Typically, blocking reagents are used to block non-specific hybridization. Such blocking reagents include, for instance, sheared and denatured salmon sperm DNA at about 100-200 μg/ml. Organic solvent, such as formamide at a concentration of about 35-50% v/v, may also be used under particular circumstances, such as for RNA:DNA hybridizations. Useful variations on these wash conditions will be readily apparent to those of ordinary skill in the art. Hybridization, particularly under high stringency conditions, may be suggestive of evolutionary similarity between the nucleotides. Such similarity is strongly indicative of a similar rôle for the nucleotides and their encoded polypeptides.

The term "hybridization complex" refers to a complex formed between two nucleic acids by

virtue of the formation of hydrogen bonds between complementary bases. A hybridization complex may be formed in solution (e.g., C_0t or R_0t analysis) or formed between one nucleic acid present in solution and another nucleic acid immobilized on a solid support (e.g., paper, membranes, filters, chips, pins or glass slides, or any other appropriate substrate to which cells or their nucleic acids have been fixed).

The words "insertion" and "addition" refer to changes in an amino acid or polynucleotide sequence resulting in the addition of one or more amino acid residues or nucleotides, respectively.

"Immune response" can refer to conditions associated with inflammation, trauma, immune disorders, or infectious or genetic disease, etc. These conditions can be characterized by expression of various factors, e.g., cytokines, chemokines, and other signaling molecules, which may affect cellular and systemic defense systems.

An "immunogenic fragment" is a polypeptide or oligopeptide fragment of CADECM which is capable of eliciting an immune response when introduced into a living organism, for example, a mammal. The term "immunogenic fragment" also includes any polypeptide or oligopeptide fragment of CADECM which is useful in any of the antibody production methods disclosed herein or known in the art.

The term "microarray" refers to an arrangement of a plurality of polynucleotides, polypeptides, antibodies, or other chemical compounds on a substrate.

The terms "element" and "array element" refer to a polynucleotide, polypeptide, antibody, or other chemical compound having a unique and defined position on a microarray.

The term "modulate" refers to a change in the activity of CADECM. For example, modulation may cause an increase or a decrease in protein activity, binding characteristics, or any other biological, functional, or immunological properties of CADECM.

The phrases "nucleic acid" and "nucleic acid sequence" refer to a nucleotide, oligonucleotide, polynucleotide, or any fragment thereof. These phrases also refer to DNA or RNA of genomic or synthetic origin which may be single-stranded or double-stranded and may represent the sense or the antisense strand, to peptide nucleic acid (PNA), or to any DNA-like or RNA-like material.

"Operably linked" refers to the situation in which a first nucleic acid sequence is placed in a functional relationship with a second nucleic acid sequence. For instance, a promoter is operably linked to a coding sequence if the promoter affects the transcription or expression of the coding sequence. Operably linked DNA sequences may be in close proximity or contiguous and, where necessary to join two protein coding regions, in the same reading frame.

"Peptide nucleic acid" (PNA) refers to an antisense molecule or anti-gene agent which

comprises an oligonucleotide of at least about 5 nucleotides in length linked to a peptide backbone of amino acid residues ending in lysine. The terminal lysine confers solubility to the composition. PNAs preferentially bind complementary single stranded DNA or RNA and stop transcript elongation, and may be pegylated to extend their lifespan in the cell.

"Post-translational modification" of an CADECM may involve lipidation, glycosylation, phosphorylation, acetylation, racemization, proteolytic cleavage, and other modifications known in the art. These processes may occur synthetically or biochemically. Biochemical modifications will vary by cell type depending on the enzymatic milieu of CADECM.

"Probe" refers to nucleic acids encoding CADECM, their complements, or fragments thereof, which are used to detect identical, allelic or related nucleic acids. Probes are isolated oligonucleotides or polynucleotides attached to a detectable label or reporter molecule. Typical labels include radioactive isotopes, ligands, chemiluminescent agents, and enzymes. "Primers" are short nucleic acids, usually DNA oligonucleotides, which may be annealed to a target polynucleotide by complementary base-pairing. The primer may then be extended along the target DNA strand by a DNA polymerase enzyme. Primer pairs can be used for amplification (and identification) of a nucleic acid, e.g., by the polymerase chain reaction (PCR).

Probes and primers as used in the present invention typically comprise at least 15 contiguous nucleotides of a known sequence. In order to enhance specificity, longer probes and primers may also be employed, such as probes and primers that comprise at least 20, 25, 30, 40, 50, 60, 70, 80, 90, 100, or at least 150 consecutive nucleotides of the disclosed nucleic acid sequences. Probes and primers may be considerably longer than these examples, and it is understood that any length supported by the specification, including the tables, figures, and Sequence Listing, may be used.

Methods for preparing and using probes and primers are described in, for example, Sambrook, J. and D.W. Russell (2001; Molecular Cloning: A Laboratory Manual, 3rd ed., vol. 1-3, Cold Spring Harbor Press, Cold Spring Harbor NY), Ausubel, F.M. et al. (1999; Short Protocols in Molecular Biology, 4th ed., John Wiley & Sons, New York NY), and Innis, M. et al. (1990; PCR Protocols, A Guide to Methods and Applications, Academic Press, San Diego CA). PCR primer pairs can be derived from a known sequence, for example, by using computer programs intended for that purpose such as Primer (Version 0.5, 1991, Whitehead Institute for Biomedical Research, Cambridge MA).

Oligonucleotides for use as primers are selected using software known in the art for such purpose. For example, OLIGO 4.06 software is useful for the selection of PCR primer pairs of up to 100 nucleotides each, and for the analysis of oligonucleotides and larger polynucleotides of up to 5,000 nucleotides from an input polynucleotide sequence of up to 32 kilobases. Similar primer selection

programs have incorporated additional features for expanded capabilities. For example, the PrimOU primer selection program (available to the public from the Genome Center at University of Texas South West Medical Center, Dallas TX) is capable of choosing specific primers from megabase sequences and is thus useful for designing primers on a genome-wide scope. The Primer3 primer selection program (available to the public from the Whitehead Institute/MIT Center for Genome Research, Cambridge MA) allows the user to input a "mispriming library," in which sequences to avoid as primer binding sites are user-specified. Primer3 is useful, in particular, for the selection of oligonucleotides for microarrays. (The source code for the latter two primer selection programs may also be obtained from their respective sources and modified to meet the user's specific needs.) The PrimeGen program (available to the public from the UK Human Genome Mapping Project Resource Centre, Cambridge UK) designs primers based on multiple sequence alignments, thereby allowing selection of primers that hybridize to either the most conserved or least conserved regions of aligned nucleic acid sequences. Hence, this program is useful for identification of both unique and conserved oligonucleotides and polynucleotide fragments. The oligonucleotides and polynucleotide fragments identified by any of the above selection methods are useful in hybridization technologies, for example, as PCR or sequencing primers, microarray elements, or specific probes to identify fully or partially complementary polynucleotides in a sample of nucleic acids. Methods of oligonucleotide selection are not limited to those described above.

A "recombinant nucleic acid" is a nucleic acid that is not naturally occurring or has a sequence that is made by an artificial combination of two or more otherwise separated segments of sequence. This artificial combination is often accomplished by chemical synthesis or, more commonly, by the artificial manipulation of isolated segments of nucleic acids, e.g., by genetic engineering techniques such as those described in Sambrook and Russell (*supra*). The term recombinant includes nucleic acids that have been altered solely by addition, substitution, or deletion of a portion of the nucleic acid. Frequently, a recombinant nucleic acid may include a nucleic acid sequence operably linked to a promoter sequence. Such a recombinant nucleic acid may be part of a vector that is used, for example, to transform a cell.

Alternatively, such recombinant nucleic acids may be part of a viral vector, e.g., based on a vaccinia virus, that could be use to vaccinate a mammal wherein the recombinant nucleic acid is expressed, inducing a protective immunological response in the mammal.

A "regulatory element" refers to a nucleic acid sequence usually derived from untranslated regions of a gene and includes enhancers, promoters, introns, and 5' and 3' untranslated regions (UTRs). Regulatory elements interact with host or viral proteins which control transcription,

translation, or RNA stability.

"Reporter molecules" are chemical or biochemical moieties used for labeling a nucleic acid, amino acid, or antibody. Reporter molecules include radionuclides; enzymes; fluorescent, chemiluminescent, or chromogenic agents; substrates; cofactors; inhibitors; magnetic particles; and other moieties known in the art.

An "RNA equivalent," in reference to a DNA molecule, is composed of the same linear sequence of nucleotides as the reference DNA molecule with the exception that all occurrences of the nitrogenous base thymine are replaced with uracil, and the sugar backbone is composed of ribose instead of deoxyribose.

The term "sample" is used in its broadest sense. A sample suspected of containing CADECM, nucleic acids encoding CADECM, or fragments thereof may comprise a bodily fluid; an extract from a cell, chromosome, organelle, or membrane isolated from a cell; a cell; genomic DNA, RNA, or cDNA, in solution or bound to a substrate; a tissue; a tissue print; etc.

The terms "specific binding" and "specifically binding" refer to that interaction between a protein or peptide and an agonist, an antibody, an antagonist, a small molecule, or any natural or synthetic binding composition. The interaction is dependent upon the presence of a particular structure of the protein, e.g., the antigenic determinant or epitope, recognized by the binding molecule. For example, if an antibody is specific for epitope "A," the presence of a polypeptide comprising the epitope A, or the presence of free unlabeled A, in a reaction containing free labeled A and the antibody will reduce the amount of labeled A that binds to the antibody.

The term "substantially purified" refers to nucleic acid or amino acid sequences that are removed from their natural environment and are isolated or separated, and are at least about 60% free, preferably at least about 75% free, and most preferably at least about 90% free from other components with which they are naturally associated.

A "substitution" refers to the replacement of one or more amino acid residues or nucleotides by different amino acid residues or nucleotides, respectively.

"Substrate" refers to any suitable rigid or semi-rigid support including membranes, filters, chips, slides, wafers, fibers, magnetic or nonmagnetic beads, gels, tubing, plates, polymers, microparticles and capillaries. The substrate can have a variety of surface forms, such as wells, trenches, pins, channels and pores, to which polynucleotides or polypeptides are bound.

A "transcript image" or "expression profile" refers to the collective pattern of gene expression by a particular cell type or tissue under given conditions at a given time.

"Transformation" describes a process by which exogenous DNA is introduced into a recipient

cell. Transformation may occur under natural or artificial conditions according to various methods well known in the art, and may rely on any known method for the insertion of foreign nucleic acid sequences into a prokaryotic or eukaryotic host cell. The method for transformation is selected based on the type of host cell being transformed and may include, but is not limited to, bacteriophage or viral infection, electroporation, heat shock, lipofection, and particle bombardment. The term "transformed cells" includes stably transformed cells in which the inserted DNA is capable of replication either as an autonomously replicating plasmid or as part of the host chromosome, as well as transiently transformed cells which express the inserted DNA or RNA for limited periods of time.

A "transgenic organism," as used herein, is any organism, including but not limited to animals and plants, in which one or more of the cells of the organism contains heterologous nucleic acid introduced by way of human intervention, such as by transgenic techniques well known in the art. The nucleic acid is introduced into the cell, directly or indirectly by introduction into a precursor of the cell, by way of deliberate genetic manipulation, such as by microinjection or by infection with a recombinant virus. In another embodiment, the nucleic acid can be introduced by infection with a recombinant viral vector, such as a lentiviral vector (Lois, C. et al. (2002) Science 295:868-872). The term genetic manipulation does not include classical cross-breeding, or *in vitro* fertilization, but rather is directed to the introduction of a recombinant DNA molecule. The transgenic organisms contemplated in accordance with the present invention include bacteria, cyanobacteria, fungi, plants and animals. The isolated DNA of the present invention can be introduced into the host by methods known in the art, for example infection, transfection, transformation or transconjugation. Techniques for transferring the DNA of the present invention into such organisms are widely known and provided in references such as Sambrook and Russell (*supra*).

A "variant" of a particular nucleic acid sequence is defined as a nucleic acid sequence having at least 40% sequence identity to the particular nucleic acid sequence over a certain length of one of the nucleic acid sequences using blastn with the "BLAST 2 Sequences" tool Version 2.0.9 (May-07-1999) set at default parameters. Such a pair of nucleic acids may show, for example, at least 50%, at least 60%, at least 70%, at least 80%, at least 85%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, or at least 99% or greater sequence identity over a certain defined length. A variant may be described as, for example, an "allelic" (as defined above), "splice," "species," or "polymorphic" variant. A splice variant may have significant identity to a reference molecule, but will generally have a greater or lesser number of polynucleotides due to alternate splicing during mRNA processing. The corresponding polypeptide may possess additional functional domains or lack domains that are present in the reference molecule.

Species variants are polynucleotides that vary from one species to another. The resulting polypeptides will generally have significant amino acid identity relative to each other. A polymorphic variant is a variation in the polynucleotide sequence of a particular gene between individuals of a given species. Polymorphic variants also may encompass "single nucleotide polymorphisms" (SNPs) in which the polynucleotide sequence varies by one nucleotide base. The presence of SNPs may be indicative of, for example, a certain population, a disease state, or a propensity for a disease state.

A "variant" of a particular polypeptide sequence is defined as a polypeptide sequence having at least 40% sequence identity or sequence similarity to the particular polypeptide sequence over a certain length of one of the polypeptide sequences using blastp with the "BLAST 2 Sequences" tool Version 2.0.9 (May-07-1999) set at default parameters. Such a pair of polypeptides may show, for example, at least 50%, at least 60%, at least 70%, at least 80%, at least 85%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, or at least 99% or greater sequence identity or sequence similarity over a certain defined length of one of the polypeptides.

THE INVENTION

Various embodiments of the invention include new human cell adhesion and extracellular matrix proteins (CADECM), the polynucleotides encoding CADECM, and the use of these compositions for the diagnosis, treatment, or prevention of immune system disorders, neurological disorders, developmental disorders, connective tissue disorders, and cell proliferative disorders, including cancer.

Table 1 summarizes the nomenclature for the full length polynucleotide and polypeptide embodiments of the invention. Each polynucleotide and its corresponding polypeptide are correlated to a single Incyte project identification number (Incyte Project ID). Each polypeptide sequence is denoted by both a polypeptide sequence identification number (Polypeptide SEQ ID NO:) and an Incyte polypeptide sequence number (Incyte Polypeptide ID) as shown. Each polynucleotide sequence is denoted by both a polynucleotide sequence identification number (Polynucleotide SEQ ID NO:) and an Incyte polynucleotide consensus sequence number (Incyte Polynucleotide ID) as shown. Column 6 shows the Incyte ID numbers of physical, full length clones corresponding to the polypeptide and polynucleotide sequences of the invention. The full length clones encode polypeptides which have at least 95% sequence identity to the polypeptide sequences shown in column 3.

Table 2 shows sequences with homology to polypeptide embodiments of the invention as identified by BLAST analysis against the GenBank protein (genpept) database and the PROTEOME

database. Columns 1 and 2 show the polypeptide sequence identification number (Polypeptide SEQ ID NO:) and the corresponding Incyte polypeptide sequence number (Incyte Polypeptide ID) for polypeptides of the invention. Column 3 shows the GenBank identification number (GenBank ID NO:) of the nearest GenBank homolog and the PROTEOME database identification numbers (PROTEOME ID NO:) of the nearest PROTEOME database homologs. Column 4 shows the probability scores for the matches between each polypeptide and its homolog(s). Column 5 shows the annotation of the GenBank and PROTEOME database homolog(s) along with relevant citations where applicable, all of which are expressly incorporated by reference herein.

Table 3 shows various structural features of the polypeptides of the invention. Columns 1 and 2 show the polypeptide sequence identification number (SEQ ID NO:) and the corresponding Incyte polypeptide sequence number (Incyte Polypeptide ID) for each polypeptide of the invention. Column 3 shows the number of amino acid residues in each polypeptide. Column 4 shows potential phosphorylation sites, and column 5 shows potential glycosylation sites, as determined by the MOTIFS program of the GCG sequence analysis software package (Accelrys, Burlington MA). Column 6 shows amino acid residues comprising signature sequences, domains, and motifs. Column 7 shows analytical methods for protein structure/function analysis and in some cases, searchable databases to which the analytical methods were applied.

Together, Tables 2 and 3 summarize the properties of polypeptides of the invention, and these properties establish that the claimed polypeptides are cell adhesion and extracellular matrix proteins. For example, SEQ ID NO:1 is 94% identical, from residue M1 to residue A908, to human nidogen-2 (GenBank ID g2791962) as determined by the Basic Local Alignment Search Tool (BLAST). (See Table 2.) The BLAST probability score is 0.0, which indicates the probability of obtaining the observed polypeptide sequence alignment by chance. SEQ ID NO:1 also has homology to proteins that are localized to the extracellular matrix, and are basement membrane proteins that bind perlecan, laminin-1, and collagens, as determined by BLAST analysis using the PROTEOME database. SEQ ID NO:1 also contains annotation of HMMER-PFAM/SMRT hit domains as determined by searching for statistically significant matches in the hidden Markov model (HMM)-based PFAM/SMART database of conserved protein families/domains. (See Table 3.) Data from BLIMPS, MOTIFS, and BLAST analyses against the PRODOM and DOMO databases, provide further corroborative evidence that SEQ ID NO:1 is a coagulation glycoprotein.

In another example, SEQ ID NO:8 is 47% identical, from residue A171 to residue G368, to mouse pro-alpha-1 type I collagen (GenBank ID g192262) as determined by BLAST. (See Table 2.) The BLAST probability score is 1.8e-45. SEQ ID NO:8 also has homology to proteins that contain

collagen triple helix repeats, and have regions of similarity to collagen type V alpha 2 and collagen typ(VI alpha 1, and may be involved in skeletal development and maintaining muscle fiber integrity, as determined by BLAST analysis using the PROTEOME database. (See Table 2.) SEQ ID NO:8 also contains a collagen triple helix repeat domain, as determined by searching for statistically significant matches in the hidden Markov model (HMM)-based PFAM database of conserved protein families/domains. (See Table 3.) Data from BLAST analyses against the PRODOM and DOMO databases provide further corroborative evidence that SEQ ID NO:8 is a collagen protein.

In another example, SEQ ID NO:10 is 100% identical, from residue M1 to residue P211, to CYR61 (GenBank ID g12584866) as determined by BLAST. (See Table 2.) The BLAST probability score is 1.6e-117. SEQ ID NO:10 also has homology to proteins that are localized to the extracellular matrix, play a role in cell adhesion, cell migration, angiogenesis and cell proliferation, and are angiogenic inducers, as determined by BLAST analysis using the PROTEOME database. SEQ ID NO:10 also contains a von Willebrand factor (vWF) type C domain as determined by searching for statistically significant matches in the hidden Markov model (HMM)-based PFAM and SMART databases of conserved protein families/domains. (See Table 3.) Data from BLIMPS, MOTIFS, and PROFILESCAN analyses, and BLAST analyses against the PRODOM and DOMO databases, provide further corroborative evidence that SEQ ID NO:10 is a cysteine-rich angiogenic inducer.

In another example, SEQ ID NO:21 is 99% identical, from residue M1 to residue E884, to protocadherin 10 (GenBank ID g13876380) as determined by BLAST. (See Table 2.) The BLAST probability score is 0.0. SEQ ID NO:21 also has homology to proteins that are localized to the plasma membrane, may play a role in the formation of neural networks through segregation of brain nuclei and mediation of axonal connnections, and are members of the cadherin subclass of calcium-dependent cell-cell adhesion molecules, as determined by BLAST analysis using the PROTEOME database. SEQ ID NO:21 also contains a cadherin domain as determined by searching for statistically significant matches in the hidden Markov model (HMM)-based PFAM and SMART databases of conserved protein families/domains. (See Table 3.) Data from BLIMPS, MOTIFS, and PROFILESCAN analyses, and BLAST analyses against the PRODOM and DOMO databases, provide further corroborative evidence that SEQ ID NO:21 is a protocadherin.

In another example, SEQ ID NO:28 is 98% identical, from residue M1 to residue D73, to human decorin (GenBank ID g181519) as determined by BLAST. (See Table 2.) The BLAST probability score is 6.4e-36. SEQ ID NO:28 also has homology to decorin, a dermatan/chondroitin sulfate proteoglycan localized to the extracellular matrix, that binds to collagen and transforming growth factor beta, and negatively controls cell growth, as determined by BLAST analysis using the

PROTEOME database. Data from BLAST analysis against the PRODOM database provides further corroborative evidence that SEQ ID NO:28 is a decorin. (See Table 3.)

In yet another example, SEQ ID NO:35 is 100% identical, from residue G171 to residue A373, to human emilin precursor (GenBank ID g5353510) as determined by BLAST. (See Table 2.) The BLAST probability score is 2.1e-210. SEQ ID NO:35 also has homology to elastin microfibril interface located protein, which is an extracellular matrix protein found between amorphous elastin and microfibrils and may play a role in elastin deposition as determined by BLAST analysis using the PROTEOME database. Further, SEQ ID NO:35 also has homology to extracellular glycoprotein EMILIN-2 precursor, which is a secreted glycoprotein, which contains a globular C1q domain, a short collagenous stalk, a coiled-coil region, a proline-rich region, and a cysteine-rich domain (EMI domain), and interacts via its gC1q domain with EMILIN as determined by BLAST analysis using the PROTEOME database. SEQ ID NO:35 also contains a complement component C1q domain and a collagen triple helix repeat (20 copies) domain as determined by searching for statistically significant matches in the hidden Markov model (HMM)-based PFAM/SMART databases of conserved protein families/domains. (See Table 3.) Data from BLIMPS and MOTIFS, and BLAST analyses against the PRODOM and DOMO databases, provide further corroborative evidence that SEQ ID NO:35 is an emilin precursor.

In a further example, SEQ ID NO:42 is 98% identical, from residue M1 to residue A237, to human acetycholinesterase collagen-like tail subunit isoform III (GenBank ID g7239359) as determined by BLAST. (See Table 2.) The BLAST probability score is 2.1e-133. SEQ ID NO:42 also has homology to proteins that are localized to the extracellular matrix, have binding function, and is a collagen-like tail subunit of asymmetric acetylcholinesterase function, as determined by BLAST analysis using the PROTEOME database. SEQ ID NO:42 also contains a collagen triple helix repeat domain as determined by searching for statistically significant matches in the hidden Markov model (HMM)-based PFAM database of conserved protein families/domains. (See Table 3.) Data from BLAST anaylses against the PRODOM and DOMO databases provide further corroborative evidence that SEQ ID NO:42 is an acetylcholinesterase. SEQ ID NO:2-7, SEQ ID NO:9, SEQ ID NO:11-20, SEQ ID NO:22-27, SEQ ID NO:29-34, and SEQ ID NO:36-41 were analyzed and annotated in a similar manner. The algorithms and parameters for the analysis of SEQ ID NO:1-42 are described in Table 7.

As shown in Table 4, the full length polynucleotide embodiments were assembled using cDNA sequences or coding (exon) sequences derived from genomic DNA, or any combination of these two types of sequences. Column 1 lists the polynucleotide sequence identification number (Polynucleotide

SEQ ID NO:), the corresponding Incyte polynucleotide consensus sequence number (Incyte ID) for each polynucleotide of the invention, and the length of each polynucleotide sequence in basepairs. Column 2 shows the nucleotide start (5') and stop (3') positions of the cDNA and/or genomic sequences used to assemble the full length polynucleotide embodiments, and of fragments of the polynucleotides which are useful, for example, in hybridization or amplification technologies that identify SEQ ID NO:43-84 or that distinguish between SEQ ID NO:43-84 and related polynucleotides.

The polynucleotide fragments described in Column 2 of Table 4 may refer specifically, for example, to Incyte cDNAs derived from tissue-specific cDNA libraries or from pooled cDNA libraries. Alternatively, the polynucleotide fragments described in column 2 may refer to GenBank cDNAs or ESTs which contributed to the assembly of the full length polynucleotides. In addition, the polynucleotide fragments described in column 2 may identify sequences derived from the ENSEMBL (The Sanger Centre, Cambridge, UK) database (*i.e.*, those sequences including the designation "ENST"). Alternatively, the polynucleotide fragments described in column 2 may be derived from the NCBI RefSeq Nucleotide Sequence Records Database (*i.e.*, those sequences including the designation "NM" or "NT") or the NCBI RefSeq Protein Sequence Records (*i.e.*, those sequences including the designation "NP"). Alternatively, the polynucleotide fragments described in column 2 may refer to assemblages of both cDNA and Genscan-predicted exons brought together by an "exon stitching" algorithm. For example, a polynucleotide sequence identified as FL_$XXXXXX$_N_1_N_2_$YYYYY$_N_3_N_4 represents a "stitched" sequence in which *XXXXXX* is the identification number of the cluster of sequences to which the algorithm was applied, and *YYYYY* is the number of the prediction generated by the algorithm, and $N_{1,2,3...}$, if present, represent specific exons that may have been manually edited during analysis (See Example V). Alternatively, the polynucleotide fragments in column 2 may refer to assemblages of exons brought together by an "exon-stretching" algorithm. For example, a polynucleotide sequence identified as FL*XXXXXX*_g*AAAAA*_g*BBBBB*_1_*N* is a "stretched" sequence, with *XXXXXX* being the Incyte project identification number, g*AAAAA* being the GenBank identification number of the human genomic sequence to which the "exon-stretching" algorithm was applied, g*BBBBB* being the GenBank identification number or NCBI RefSeq identification number of the nearest GenBank protein homolog, and *N* referring to specific exons (See Example V). In instances where a RefSeq sequence was used as a protein homolog for the "exon-stretching" algorithm, a RefSeq identifier (denoted by "NM," "NP," or "NT") may be used in place of the GenBank identifier (*i.e.*, g*BBBBB*).

Alternatively, a prefix identifies component sequences that were hand-edited, predicted from genomic DNA sequences, or derived from a combination of sequence analysis methods. The

following Table lists examples of component sequence prefixes and corresponding sequence analysis methods associated with the prefixes (see Example IV and Example V).

Prefix	Type of analysis and/or examples of programs
GNN, GFG, ENST	Exon prediction from genomic sequences using, for example, GENSCAN (Stanford University, CA, USA) or FGENES (Computer Genomics Group, The Sanger Centre, Cambridge, UK).
GBI	Hand-edited analysis of genomic sequences.
FL	Stitched or stretched genomic sequences (see Example V).
INCY	Full length transcript and exon prediction from mapping of EST sequences to the genome. Genomic location and EST composition data are combined to predict the exons and resulting transcript.

In some cases, Incyte cDNA coverage redundant with the sequence coverage shown in Table 4 was obtained to confirm the final consensus polynucleotide sequence, but the relevant Incyte cDNA identification numbers are not shown.

Table 5 shows the representative cDNA libraries for those full length polynucleotides which were assembled using Incyte cDNA sequences. The representative cDNA library is the Incyte cDNA library which is most frequently represented by the Incyte cDNA sequences which were used to assemble and confirm the above polynucleotides. The tissues and vectors which were used to construct the cDNA libraries shown in Table 5 are described in Table 6.

Table 8 shows single nucleotide polymorphisms (SNPs) found in polynucleotide sequences of the invention, along with allele frequencies in different human populations. Columns 1 and 2 show the polynucleotide sequence identification number (SEQ ID NO:) and the corresponding Incyte project identification number (PID) for polynucleotides of the invention. Column 3 shows the Incyte identification number for the EST in which the SNP was detected (EST ID), and column 4 shows the identification number for the SNP (SNP ID). Column 5 shows the position within the EST sequence at which the SNP is located (EST SNP), and column 6 shows the position of the SNP within the full-length polynucleotide sequence (CB1 SNP). Column 7 shows the allele found in the EST sequence. Columns 8 and 9 show the two alleles found at the SNP site. Column 10 shows the amino acid encoded by the codon including the SNP site, based upon the allele found in the EST. Columns 11-14 show the frequency of allele 1 in four different human populations. An entry of n/d (not detected) indicates that the frequency of allele 1 in the population was too low to be detected, while n/a (not available) indicates that the allele frequency was not determined for the population.

The invention also encompasses CADECM variants. Various embodiments of CADECM variants can have at least about 80%, at least about 90%, or at least about 95% amino acid sequence identity to the CADECM amino acid sequence, and can contain at least one functional or structural characteristic of CADECM.

Various embodiments also encompass polynucleotides which encode CADECM. In a particular embodiment, the invention encompasses a polynucleotide sequence comprising a sequence selected from the group consisting of SEQ ID NO:43-84, which encodes CADECM. The polynucleotide sequences of SEQ ID NO:43-84, as presented in the Sequence Listing, embrace the equivalent RNA sequences, wherein occurrences of the nitrogenous base thymine are replaced with uracil, and the sugar backbone is composed of ribose instead of deoxyribose.

The invention also encompasses variants of a polynucleotide encoding CADECM. In particular, such a variant polynucleotide will have at least about 70%, or alternatively at least about 85%, or even at least about 95% polynucleotide sequence identity to a polynucleotide encoding CADECM. A particular aspect of the invention encompasses a variant of a polynucleotide comprising a sequence selected from the group consisting of SEQ ID NO:43-84 which has at least about 70%, or alternatively at least about 85%, or even at least about 95% polynucleotide sequence identity to a nucleic acid sequence selected from the group consisting of SEQ ID NO:43-84. Any one of the polynucleotide variants described above can encode a polypeptide which contains at least one functional or structural characteristic of CADECM.

In addition, or in the alternative, a polynucleotide variant of the invention is a splice variant of a polynucleotide encoding CADECM. A splice variant may have portions which have significant sequence identity to a polynucleotide encoding CADECM, but will generally have a greater or lesser number of nucleotides due to additions or deletions of blocks of sequence arising from alternate splicing during mRNA processing. A splice variant may have less than about 70%, or alternatively less than about 60%, or alternatively less than about 50% polynucleotide sequence identity to a polynucleotide encoding CADECM over its entire length; however, portions of the splice variant will have at least about 70%, or alternatively at least about 85%, or alternatively at least about 95%, or alternatively 100% polynucleotide sequence identity to portions of the polynucleotide encoding CADECM. For example, a polynucleotide comprising a sequence of SEQ ID NO:51 and a polynucleotide comprising a sequence of SEQ ID NO:71 are splice variants of each other; a polynucleotide comprising a sequence of SEQ ID NO:55 and a polynucleotide comprising a sequence of SEQ ID NO:56 are splice variants of each other; a polynucleotide comprising a sequence of SEQ ID NO:58 and a polynucleotide comprising a sequence of SEQ ID NO:59 are splice variants of each

other; a polynucleotide comprising a sequence of SEQ ID NO:69 and a polynucleotide comprising a sequence of SEQ ID NO:73 are splice variants of each other; and a polynucleotide comprising a sequence of SEQ ID NO:67, a polynucleotide comprising a sequence of SEQ ID NO:68 and a polynucleotide comprising a sequence of SEQ ID NO:84 are splice variants of each other. Any one of the splice variants described above can encode a polypeptide which contains at least one functional or structural characteristic of CADECM.

It will be appreciated by those skilled in the art that as a result of the degeneracy of the genetic code, a multitude of polynucleotide sequences encoding CADECM, some bearing minimal similarity to the polynucleotide sequences of any known and naturally occurring gene, may be produced. Thus, the invention contemplates each and every possible variation of polynucleotide sequence that could be made by selecting combinations based on possible codon choices. These combinations are made in accordance with the standard triplet genetic code as applied to the polynucleotide sequence of naturally occurring CADECM, and all such variations are to be considered as being specifically disclosed.

Although polynucleotides which encode CADECM and its variants are generally capable of hybridizing to polynucleotides encoding naturally occurring CADECM under appropriately selected conditions of stringency, it may be advantageous to produce polynucleotides encoding CADECM or its derivatives possessing a substantially different codon usage, e.g., inclusion of non-naturally occurring codons. Codons may be selected to increase the rate at which expression of the peptide occurs in a particular prokaryotic or eukaryotic host in accordance with the frequency with which particular codons are utilized by the host. Other reasons for substantially altering the nucleotide sequence encoding CADECM and its derivatives without altering the encoded amino acid sequences include the production of RNA transcripts having more desirable properties, such as a greater half-life, than transcripts produced from the naturally occurring sequence.

The invention also encompasses production of polynucleotides which encode CADECM and CADECM derivatives, or fragments thereof, entirely by synthetic chemistry. After production, the synthetic polynucleotide may be inserted into any of the many available expression vectors and cell systems using reagents well known in the art. Moreover, synthetic chemistry may be used to introduce mutations into a polynucleotide encoding CADECM or any fragment thereof.

Embodiments of the invention can also include polynucleotides that are capable of hybridizing to the claimed polynucleotides, and, in particular, to those having the sequences shown in SEQ ID NO:43-84 and fragments thereof, under various conditions of stringency (Wahl, G.M. and S.L. Berger (1987) Methods Enzymol. 152:399-407; Kimmel, A.R. (1987) Methods Enzymol. 152:507-511).

Hybridization conditions, including annealing and wash conditions, are described in "Definitions."

Methods for DNA sequencing are well known in the art and may be used to practice any of the embodiments of the invention. The methods may employ such enzymes as the Klenow fragment of DNA polymerase I, SEQUENASE (US Biochemical, Cleveland OH), Taq polymerase (Applied Biosystems), thermostable T7 polymerase (Amersham Biosciences, Piscataway NJ), or combinations of polymerases and proofreading exonucleases such as those found in the ELONGASE amplification system (Invitrogen, Carlsbad CA). Preferably, sequence preparation is automated with machines such as the MICROLAB 2200 liquid transfer system (Hamilton, Reno NV), PTC200 thermal cycler (MJ Research, Watertown MA) and ABI CATALYST 800 thermal cycler (Applied Biosystems). Sequencing is then carried out using either the ABI 373 or 377 DNA sequencing system (Applied Biosystems), the MEGABACE 1000 DNA sequencing system (Amersham Biosciences), or other systems known in the art. The resulting sequences are analyzed using a variety of algorithms which are well known in the art (Ausubel et al., *supra*, ch. 7; Meyers, R.A. (1995) Molecular Biology and Biotechnology, Wiley VCH, New York NY, pp. 856-853).

The nucleic acids encoding CADECM may be extended utilizing a partial nucleotide sequence and employing various PCR-based methods known in the art to detect upstream sequences, such as promoters and regulatory elements. For example, one method which may be employed, restriction-site PCR, uses universal and nested primers to amplify unknown sequence from genomic DNA within a cloning vector (Sarkar, G. (1993) PCR Methods Applic. 2:318-322). Another method, inverse PCR, uses primers that extend in divergent directions to amplify unknown sequence from a circularized template. The template is derived from restriction fragments comprising a known genomic locus and surrounding sequences (Triglia, T. et al. (1988) Nucleic Acids Res. 16:8186). A third method, capture PCR, involves PCR amplification of DNA fragments adjacent to known sequences in human and yeast artificial chromosome DNA (Lagerstrom, M. et al. (1991) PCR Methods Applic. 1:111-119). In this method, multiple restriction enzyme digestions and ligations may be used to insert an engineered double-stranded sequence into a region of unknown sequence before performing PCR. Other methods which may be used to retrieve unknown sequences are known in the art (Parker, J.D. et al. (1991) Nucleic Acids Res. 19:3055-3060). Additionally, one may use PCR, nested primers, and PROMOTERFINDER libraries (BD Clontech, Palo Alto CA) to walk genomic DNA. This procedure avoids the need to screen libraries and is useful in finding intron/exon junctions. For all PCR-based methods, primers may be designed using commercially available software, such as OLIGO 4.06 primer analysis software (National Biosciences, Plymouth MN) or another appropriate program, to be about 22 to 30 nucleotides in length, to have a GC content of about 50% or more, and

to anneal to the template at temperatures of about 68°C to 72°C.

When screening for full length cDNAs, it is preferable to use libraries that have been size-selected to include larger cDNAs. In addition, random-primed libraries, which often include sequences containing the 5' regions of genes, are preferable for situations in which an oligo d(T) library does not yield a full-length cDNA. Genomic libraries may be useful for extension of sequence into 5' non-transcribed regulatory regions.

Capillary electrophoresis systems which are commercially available may be used to analyze the size or confirm the nucleotide sequence of sequencing or PCR products. In particular, capillary sequencing may employ flowable polymers for electrophoretic separation, four different nucleotide-specific, laser-stimulated fluorescent dyes, and a charge coupled device camera for detection of the emitted wavelengths. Output/light intensity may be converted to electrical signal using appropriate software (e.g., GENOTYPER and SEQUENCE NAVIGATOR, Applied Biosystems), and the entire process from loading of samples to computer analysis and electronic data display may be computer controlled. Capillary electrophoresis is especially preferable for sequencing small DNA fragments which may be present in limited amounts in a particular sample.

In another embodiment of the invention, polynucleotides or fragments thereof which encode CADECM may be cloned in recombinant DNA molecules that direct expression of CADECM, or fragments or functional equivalents thereof, in appropriate host cells. Due to the inherent degeneracy of the genetic code, other polynucleotides which encode substantially the same or a functionally equivalent polypeptides may be produced and used to express CADECM.

The polynucleotides of the invention can be engineered using methods generally known in the art in order to alter CADECM-encoding sequences for a variety of purposes including, but not limited to, modification of the cloning, processing, and/or expression of the gene product. DNA shuffling by random fragmentation and PCR reassembly of gene fragments and synthetic oligonucleotides may be used to engineer the nucleotide sequences. For example, oligonucleotide-mediated site-directed mutagenesis may be used to introduce mutations that create new restriction sites, alter glycosylation patterns, change codon preference, produce splice variants, and so forth.

The nucleotides of the present invention may be subjected to DNA shuffling techniques such as MOLECULARBREEDING (Maxygen Inc., Santa Clara CA; described in U.S. Patent No. 5,837,458; Chang, C.-C. et al. (1999) Nat. Biotechnol. 17:793-797; Christians, F.C. et al. (1999) Nat. Biotechnol. 17:259-264; and Crameri, A. et al. (1996) Nat. Biotechnol. 14:315-319) to alter or improve the biological properties of CADECM, such as its biological or enzymatic activity or its ability to bind to other molecules or compounds. DNA shuffling is a process by which a library of gene variants is

produced using PCR-mediated recombination of gene fragments. The library is then subjected to selection or screening procedures that identify those gene variants with the desired properties. These preferred variants may then be pooled and further subjected to recursive rounds of DNA shuffling and selection/screening. Thus, genetic diversity is created through "artificial" breeding and rapid molecular evolution. For example, fragments of a single gene containing random point mutations may be recombined, screened, and then reshuffled until the desired properties are optimized. Alternatively, fragments of a given gene may be recombined with fragments of homologous genes in the same gene family, either from the same or different species, thereby maximizing the genetic diversity of multiple naturally occurring genes in a directed and controllable manner.

In another embodiment, polynucleotides encoding CADECM may be synthesized, in whole or in part, using one or more chemical methods well known in the art (Caruthers, M.H. et al. (1980) Nucleic Acids Symp. Ser. 7:215-223; Horn, T. et al. (1980) Nucleic Acids Symp. Ser. 7:225-232). Alternatively, CADECM itself or a fragment thereof may be synthesized using chemical methods known in the art. For example, peptide synthesis can be performed using various solution-phase or solid-phase techniques (Creighton, T. (1984) Proteins, Structures and Molecular Properties, WH Freeman, New York NY, pp. 55-60; Roberge, J.Y. et al. (1995) Science 269:202-204). Automated synthesis may be achieved using the ABI 431A peptide synthesizer (Applied Biosystems). Additionally, the amino acid sequence of CADECM, or any part thereof, may be altered during direct synthesis and/or combined with sequences from other proteins, or any part thereof, to produce a variant polypeptide or a polypeptide having a sequence of a naturally occurring polypeptide.

The peptide may be substantially purified by preparative high performance liquid chromatography (Chiez, R.M. and F.Z. Regnier (1990) Methods Enzymol. 182:392-421). The composition of the synthetic peptides may be confirmed by amino acid analysis or by sequencing (Creighton, *supra*, pp. 28-53).

In order to express a biologically active CADECM, the polynucleotides encoding CADECM or derivatives thereof may be inserted into an appropriate expression vector, i.e., a vector which contains the necessary elements for transcriptional and translational control of the inserted coding sequence in a suitable host. These elements include regulatory sequences, such as enhancers, constitutive and inducible promoters, and 5' and 3' untranslated regions in the vector and in polynucleotides encoding CADECM. Such elements may vary in their strength and specificity. Specific initiation signals may also be used to achieve more efficient translation of polynucleotides encoding CADECM. Such signals include the ATG initiation codon and adjacent sequences, e.g. the Kozak sequence. In cases where a polynucleotide sequence encoding CADECM and its initiation

codon and upstream regulatory sequences are inserted into the appropriate expression vector, no additional transcriptional or translational control signals may be needed. However, in cases where only coding sequence, or a fragment thereof, is inserted, exogenous translational control signals including an in-frame ATG initiation codon should be provided by the vector. Exogenous translational elements and initiation codons may be of various origins, both natural and synthetic. The efficiency of expression may be enhanced by the inclusion of enhancers appropriate for the particular host cell system used (Scharf, D. et al. (1994) Results Probl. Cell Differ. 20:125-162).

Methods which are well known to those skilled in the art may be used to construct expression vectors containing polynucleotides encoding CADECM and appropriate transcriptional and translational control elements. These methods include *in vitro* recombinant DNA techniques, synthetic techniques, and *in vivo* genetic recombination (Sambrook and Russell, *supra*, ch. 1-4, and 8; Ausubel et al., *supra*, ch. 1, 3, and 15).

A variety of expression vector/host systems may be utilized to contain and express polynucleotides encoding CADECM. These include, but are not limited to, microorganisms such as bacteria transformed with recombinant bacteriophage, plasmid, or cosmid DNA expression vectors; yeast transformed with yeast expression vectors; insect cell systems infected with viral expression vectors (e.g., baculovirus); plant cell systems transformed with viral expression vectors (e.g., cauliflower mosaic virus, CaMV, or tobacco mosaic virus, TMV) or with bacterial expression vectors (e.g., Ti or pBR322 plasmids); or animal cell systems (Sambrook and Russell, *supra*; Ausubel et al., *supra*; Van Heeke, G. and S.M. Schuster (1989) J. Biol. Chem. 264:5503-5509; Engelhard, E.K. et al. (1994) Proc. Natl. Acad. Sci. USA 91:3224-3227; Sandig, V. et al. (1996) Hum. Gene Ther. 7:1937-1945; Takamatsu, N. (1987) EMBO J. 6:307-311; The McGraw Hill Yearbook of Science and Technology (1992) McGraw Hill, New York NY, pp. 191-196; Logan, J. and T. Shenk (1984) Proc. Natl. Acad. Sci. USA 81:3655-3659; Harrington, J.J. et al. (1997) Nat. Genet. 15:345-355). Expression vectors derived from retroviruses, adenoviruses, or herpes or vaccinia viruses, or from various bacterial plasmids, may be used for delivery of polynucleotides to the targeted organ, tissue, or cell population (Di Nicola, M. et al. (1998) Cancer Gen. Ther. 5:350-356; Yu, M. et al. (1993) Proc. Natl. Acad. Sci. USA 90:6340-6344; Buller, R.M. et al. (1985) Nature 317:813-815; McGregor, D.P. et al. (1994) Mol. Immunol. 31:219-226; Verma, I.M. and N. Somia (1997) Nature 389:239-242). The invention is not limited by the host cell employed.

In bacterial systems, a number of cloning and expression vectors may be selected depending upon the use intended for polynucleotides encoding CADECM. For example, routine cloning, subcloning, and propagation of polynucleotides encoding CADECM can be achieved using a

multifunctional *E. coli* vector such as PBLUESCRIPT (Stratagene, La Jolla CA) or PSPORT1 plasmid (Invitrogen). Ligation of polynucleotides encoding CADECM into the vector's multiple cloning site disrupts the *lacZ* gene, allowing a colorimetric screening procedure for identification of transformed bacteria containing recombinant molecules. In addition, these vectors may be useful for *in vitro* transcription, dideoxy sequencing, single strand rescue with helper phage, and creation of nested deletions in the cloned sequence (Van Heeke, G. and S.M. Schuster (1989) J. Biol. Chem. 264:5503-5509). When large quantities of CADECM are needed, e.g. for the production of antibodies, vectors which direct high level expression of CADECM may be used. For example, vectors containing the strong, inducible SP6 or T7 bacteriophage promoter may be used.

Yeast expression systems may be used for production of CADECM. A number of vectors containing constitutive or inducible promoters, such as alpha factor, alcohol oxidase, and PGH promoters, may be used in the yeast *Saccharomyces cerevisiae* or *Pichia pastoris*. In addition, such vectors direct either the secretion or intracellular retention of expressed proteins and enable integration of foreign polynucleotide sequences into the host genome for stable propagation (Ausubel et al., *supra*; Bitter, G.A. et al. (1987) Methods Enzymol. 153:516-544; Scorer, C.A. et al. (1994) Bio/Technology 12:181-184).

Plant systems may also be used for expression of CADECM. Transcription of polynucleotides encoding CADECM may be driven by viral promoters, e.g., the 35S and 19S promoters of CaMV used alone or in combination with the omega leader sequence from TMV (Takamatsu, N. (1987) EMBO J. 6:307-311). Alternatively, plant promoters such as the small subunit of RUBISCO or heat shock promoters may be used (Coruzzi, G. et al. (1984) EMBO J. 3:1671-1680; Broglie, R. et al. (1984) Science 224:838-843; Winter, J. et al. (1991) Results Probl. Cell Differ. 17:85-105). These constructs can be introduced into plant cells by direct DNA transformation or pathogen-mediated transfection (The McGraw Hill Yearbook of Science and Technology (1992) McGraw Hill, New York NY, pp. 191-196).

In mammalian cells, a number of viral-based expression systems may be utilized. In cases where an adenovirus is used as an expression vector, polynucleotides encoding CADECM may be ligated into an adenovirus transcription/translation complex consisting of the late promoter and tripartite leader sequence. Insertion in a non-essential E1 or E3 region of the viral genome may be used to obtain infective virus which expresses CADECM in host cells (Logan, J. and T. Shenk (1984) Proc. Natl. Acad. Sci. USA 81:3655-3659). In addition, transcription enhancers, such as the Rous sarcoma virus (RSV) enhancer, may be used to increase expression in mammalian host cells. SV40 or EBV-based vectors may also be used for high-level protein expression.

Human artificial chromosomes (HACs) may also be employed to deliver larger fragments of DNA than can be contained in and expressed from a plasmid. HACs of about 6 kb to 10 Mb are constructed and delivered via conventional delivery methods (liposomes, polycationic amino polymers, or vesicles) for therapeutic purposes (Harrington, J.J. et al. (1997) Nat. Genet. 15:345-355).

For long term production of recombinant proteins in mammalian systems, stable expression of CADECM in cell lines is preferred. For example, polynucleotides encoding CADECM can be transformed into cell lines using expression vectors which may contain viral origins of replication and/or endogenous expression elements and a selectable marker gene on the same or on a separate vector. Following the introduction of the vector, cells may be allowed to grow for about 1 to 2 days in enriched media before being switched to selective media. The purpose of the selectable marker is to confer resistance to a selective agent, and its presence allows growth and recovery of cells which successfully express the introduced sequences. Resistant clones of stably transformed cells may be propagated using tissue culture techniques appropriate to the cell type.

Any number of selection systems may be used to recover transformed cell lines. These include, but are not limited to, the herpes simplex virus thymidine kinase and adenine phosphoribosyltransferase genes, for use in *tk*$^-$ and *apr*$^-$ cells, respectively (Wigler, M. et al. (1977) Cell 11:223-232; Lowy, I. et al. (1980) Cell 22:817-823). Also, antimetabolite, antibiotic, or herbicide resistance can be used as the basis for selection. For example, *dhfr* confers resistance to methotrexate; *neo* confers resistance to the aminoglycosides neomycin and G-418; and *als* and *pat* confer resistance to chlorsulfuron and phosphinotricin acetyltransferase, respectively (Wigler, M. et al. (1980) Proc. Natl. Acad. Sci. USA 77:3567-3570; Colbere-Garapin, F. et al. (1981) J. Mol. Biol. 150:1-14). Additional selectable genes have been described, e.g., *trpB* and *hisD*, which alter cellular requirements for metabolites (Hartman, S.C. and R.C. Mulligan (1988) Proc. Natl. Acad. Sci. USA 85:8047-8051). Visible markers, e.g., anthocyanins, green fluorescent proteins (GFP; BD Clontech), β-glucuronidase and its substrate β-glucuronide, or luciferase and its substrate luciferin may be used. These markers can be used not only to identify transformants, but also to quantify the amount of transient or stable protein expression attributable to a specific vector system (Rhodes, C.A. (1995) Methods Mol. Biol. 55:121-131).

Although the presence/absence of marker gene expression suggests that the gene of interest is also present, the presence and expression of the gene may need to be confirmed. For example, if the sequence encoding CADECM is inserted within a marker gene sequence, transformed cells containing polynucleotides encoding CADECM can be identified by the absence of marker gene function. Alternatively, a marker gene can be placed in tandem with a sequence encoding CADECM

under the control of a single promoter. Expression of the marker gene in response to induction or selection usually indicates expression of the tandem gene as well.

In general, host cells that contain the polynucleotide encoding CADECM and that express CADECM may be identified by a variety of procedures known to those of skill in the art. These procedures include, but are not limited to, DNA-DNA or DNA-RNA hybridizations, PCR amplification, and protein bioassay or immunoassay techniques which include membrane, solution, or chip based technologies for the detection and/or quantification of nucleic acid or protein sequences.

Immunological methods for detecting and measuring the expression of CADECM using either specific polyclonal or monoclonal antibodies are known in the art. Examples of such techniques include enzyme-linked immunosorbent assays (ELISAs), radioimmunoassays (RIAs), and fluorescence activated cell sorting (FACS). A two-site, monoclonal-based immunoassay utilizing monoclonal antibodies reactive to two non-interfering epitopes on CADECM is preferred, but a competitive binding assay may be employed. These and other assays are well known in the art (Hampton, R. et al. (1990) Serological Methods, a Laboratory Manual, APS Press, St. Paul MN, Sect. IV; Coligan, J.E. et al. (1997) Current Protocols in Immunology, Greene Pub. Associates and Wiley-Interscience, New York NY; Pound, J.D. (1998) Immunochemical Protocols, Humana Press, Totowa NJ).

A wide variety of labels and conjugation techniques are known by those skilled in the art and may be used in various nucleic acid and amino acid assays. Means for producing labeled hybridization or PCR probes for detecting sequences related to polynucleotides encoding CADECM include oligolabeling, nick translation, end-labeling, or PCR amplification using a labeled nucleotide. Alternatively, polynucleotides encoding CADECM, or any fragments thereof, may be cloned into a vector for the production of an mRNA probe. Such vectors are known in the art, are commercially available, and may be used to synthesize RNA probes *in vitro* by addition of an appropriate RNA polymerase such as T7, T3, or SP6 and labeled nucleotides. These procedures may be conducted using a variety of commercially available kits, such as those provided by Amersham Biosciences, Promega (Madison WI), and US Biochemical. Suitable reporter molecules or labels which may be used for ease of detection include radionuclides, enzymes, fluorescent, chemiluminescent, or chromogenic agents, as well as substrates, cofactors, inhibitors, magnetic particles, and the like.

Host cells transformed with polynucleotides encoding CADECM may be cultured under conditions suitable for the expression and recovery of the protein from cell culture. The protein produced by a transformed cell may be secreted or retained intracellularly depending on the sequence and/or the vector used. As will be understood by those of skill in the art, expression vectors containing

polynucleotides which encode CADECM may be designed to contain signal sequences which direct secretion of CADECM through a prokaryotic or eukaryotic cell membrane.

In addition, a host cell strain may be chosen for its ability to modulate expression of the inserted polynucleotides or to process the expressed protein in the desired fashion. Such modifications of the polypeptide include, but are not limited to, acetylation, carboxylation, glycosylation, phosphorylation, lipidation, and acylation. Post-translational processing which cleaves a "prepro" or "pro" form of the protein may also be used to specify protein targeting, folding, and/or activity. Different host cells which have specific cellular machinery and characteristic mechanisms for post-translational activities (e.g., CHO, HeLa, MDCK, HEK293, and WI38) are available from the American Type Culture Collection (ATCC, Manassas VA) and may be chosen to ensure the correct modification and processing of the foreign protein.

In another embodiment of the invention, natural, modified, or recombinant polynucleotides encoding CADECM may be ligated to a heterologous sequence resulting in translation of a fusion protein in any of the aforementioned host systems. For example, a chimeric CADECM protein containing a heterologous moiety that can be recognized by a commercially available antibody may facilitate the screening of peptide libraries for inhibitors of CADECM activity. Heterologous protein and peptide moieties may also facilitate purification of fusion proteins using commercially available affinity matrices. Such moieties include, but are not limited to, glutathione S-transferase (GST), maltose binding protein (MBP), thioredoxin (Trx), calmodulin binding peptide (CBP), 6-His, FLAG, *c-myc*, and hemagglutinin (HA). GST, MBP, Trx, CBP, and 6-His enable purification of their cognate fusion proteins on immobilized glutathione, maltose, phenylarsine oxide, calmodulin, and metal-chelate resins, respectively. FLAG, *c-myc*, and hemagglutinin (HA) enable immunoaffinity purification of fusion proteins using commercially available monoclonal and polyclonal antibodies that specifically recognize these epitope tags. A fusion protein may also be engineered to contain a proteolytic cleavage site located between the CADECM encoding sequence and the heterologous protein sequence, so that CADECM may be cleaved away from the heterologous moiety following purification. Methods for fusion protein expression and purification are discussed in Ausubel et al. (*supra*, ch. 10 and 16). A variety of commercially available kits may also be used to facilitate expression and purification of fusion proteins.

In another embodiment, synthesis of radiolabeled CADECM may be achieved *in vitro* using the TNT rabbit reticulocyte lysate or wheat germ extract system (Promega). These systems couple transcription and translation of protein-coding sequences operably associated with the T7, T3, or SP6 promoters. Translation takes place in the presence of a radiolabeled amino acid precursor, for

example, ^{35}S-methionine.

CADECM, fragments of CADECM, or variants of CADECM may be used to screen for compounds that specifically bind to CADECM. One or more test compounds may be screened for specific binding to CADECM. In various embodiments, 1, 2, 3, 4, 5, 10, 20, 50, 100, or 200 test compounds can be screened for specific binding to CADECM. Examples of test compounds can include antibodies, anticalins, oligonucleotides, proteins (e.g., ligands or receptors), or small molecules.

In related embodiments, variants of CADECM can be used to screen for binding of test compounds, such as antibodies, to CADECM, a variant of CADECM, or a combination of CADECM and/or one or more variants CADECM. In an embodiment, a variant of CADECM can be used to screen for compounds that bind to a variant of CADECM, but not to CADECM having the exact sequence of a sequence of SEQ ID NO:1-42. CADECM variants used to perform such screening can have a range of about 50% to about 99% sequence identity to CADECM, with various embodiments having 60%, 70%, 75%, 80%, 85%, 90%, and 95% sequence identity.

In an embodiment, a compound identified in a screen for specific binding to CADECM can be closely related to the natural ligand of CADECM, e.g., a ligand or fragment thereof, a natural substrate, a structural or functional mimetic, or a natural binding partner (Coligan, J.E. et al. (1991) Current Protocols in Immunology 1(2):Chapter 5). In another embodiment, the compound thus identified can be a natural ligand of a receptor CADECM (Howard, A.D. et al. (2001) Trends Pharmacol. Sci.22:132-140; Wise, A. et al. (2002) Drug Discovery Today 7:235-246).

In other embodiments, a compound identified in a screen for specific binding to CADECM can be closely related to the natural receptor to which CADECM binds, at least a fragment of the receptor, or a fragment of the receptor including all or a portion of the ligand binding site or binding pocket. For example, the compound may be a receptor for CADECM which is capable of propagating a signal, or a decoy receptor for CADECM which is not capable of propagating a signal (Ashkenazi, A. and V.M. Divit (1999) Curr. Opin. Cell Biol. 11:255-260; Mantovani, A. et al. (2001) Trends Immunol. 22:328-336). The compound can be rationally designed using known techniques. Examples of such techniques include those used to construct the compound etanercept (ENBREL; Amgen Inc., Thousand Oaks CA), which is efficacious for treating rheumatoid arthritis in humans. Etanercept is an engineered p75 tumor necrosis factor (TNF) receptor dimer linked to the Fc portion of human IgG_1 (Taylor, P.C. et al. (2001) Curr. Opin. Immunol. 13:611-616).

In one embodiment, two or more antibodies having similar or, alternatively, different specificities can be screened for specific binding to CADECM, fragments of CADECM, or variants of CADECM. The binding specificity of the antibodies thus screened can thereby be selected to

identify particular fragments or variants of CADECM. In one embodiment, an antibody can be selected such that its binding specificity allows for preferential identification of specific fragments or variants of CADECM. In another embodiment, an antibody can be selected such that its binding specificity allows for preferential diagnosis of a specific disease or condition having increased, decreased, or otherwise abnormal production of CADECM.

In an embodiment, anticalins can be screened for specific binding to CADECM, fragments of CADECM, or variants of CADECM. Anticalins are ligand-binding proteins that have been constructed based on a lipocalin scaffold (Weiss, G.A. and H.B. Lowman (2000) Chem. Biol. 7:R177-R184; Skerra, A. (2001) J. Biotechnol. 74:257-275). The protein architecture of lipocalins can include a beta-barrel having eight antiparallel beta-strands, which supports four loops at its open end. These loops form the natural ligand-binding site of the lipocalins, a site which can be re-engineered *in vitro* by amino acid substitutions to impart novel binding specificities. The amino acid substitutions can be made using methods known in the art or described herein, and can include conservative substitutions (e.g., substitutions that do not alter binding specificity) or substitutions that modestly, moderately, or significantly alter binding specificity.

In one embodiment, screening for compounds which specifically bind to, stimulate, or inhibit CADECM involves producing appropriate cells which express CADECM, either as a secreted protein or on the cell membrane. Preferred cells can include cells from mammals, yeast, *Drosophila*, or *E. coli*. Cells expressing CADECM or cell membrane fractions which contain CADECM are then contacted with a test compound and binding, stimulation, or inhibition of activity of either CADECM or the compound is analyzed.

An assay may simply test binding of a test compound to the polypeptide, wherein binding is detected by a fluorophore, radioisotope, enzyme conjugate, or other detectable label. For example, the assay may comprise the steps of combining at least one test compound with CADECM, either in solution or affixed to a solid support, and detecting the binding of CADECM to the compound. Alternatively, the assay may detect or measure binding of a test compound in the presence of a labeled competitor. Additionally, the assay may be carried out using cell-free preparations, chemical libraries, or natural product mixtures, and the test compound(s) may be free in solution or affixed to a solid support.

An assay can be used to assess the ability of a compound to bind to its natural ligand and/or to inhibit the binding of its natural ligand to its natural receptors. Examples of such assays include radio-labeling assays such as those described in U.S. Patent No. 5,914,236 and U.S. Patent No. 6,372,724. In a related embodiment, one or more amino acid substitutions can be introduced into a polypeptide

compound (such as a receptor) to improve or alter its ability to bind to its natural ligands (Matthews, D.J. and J.A. Wells. (1994) Chem. Biol. 1:25-30). In another related embodiment, one or more amino acid substitutions can be introduced into a polypeptide compound (such as a ligand) to improve or alter its ability to bind to its natural receptors (Cunningham, B.C. and J.A. Wells (1991) Proc. Natl. Acad. Sci. USA 88:3407-3411; Lowman, H.B. et al. (1991) J. Biol. Chem. 266:10982-10988).

CADECM, fragments of CADECM, or variants of CADECM may be used to screen for compounds that modulate the activity of CADECM. Such compounds may include agonists, antagonists, or partial or inverse agonists. In one embodiment, an assay is performed under conditions permissive for CADECM activity, wherein CADECM is combined with at least one test compound, and the activity of CADECM in the presence of a test compound is compared with the activity of CADECM in the absence of the test compound. A change in the activity of CADECM in the presence of the test compound is indicative of a compound that modulates the activity of CADECM. Alternatively, a test compound is combined with an *in vitro* or cell-free system comprising CADECM under conditions suitable for CADECM activity, and the assay is performed. In either of these assays, a test compound which modulates the activity of CADECM may do so indirectly and need not come in direct contact with the test compound. At least one and up to a plurality of test compounds may be screened.

In another embodiment, polynucleotides encoding CADECM or their mammalian homologs may be "knocked out" in an animal model system using homologous recombination in embryonic stem (ES) cells. Such techniques are well known in the art and are useful for the generation of animal models of human disease (see, e.g., U.S. Patent No. 5,175,383 and U.S. Patent No. 5,767,337). For example, mouse ES cells, such as the mouse 129/SvJ cell line, are derived from the early mouse embryo and grown in culture. The ES cells are transformed with a vector containing the gene of interest disrupted by a marker gene, e.g., the neomycin phosphotransferase gene (*neo*; Capecchi, M.R. (1989) Science 244:1288-1292). The vector integrates into the corresponding region of the host genome by homologous recombination. Alternatively, homologous recombination takes place using the Cre-loxP system to knockout a gene of interest in a tissue- or developmental stage-specific manner (Marth, J.D. (1996) Clin. Invest. 97:1999-2002; Wagner, K.U. et al. (1997) Nucleic Acids Res. 25:4323-4330). Transformed ES cells are identified and microinjected into mouse cell blastocysts such as those from the C57BL/6 mouse strain. The blastocysts are surgically transferred to pseudopregnant dams, and the resulting chimeric progeny are genotyped and bred to produce heterozygous or homozygous strains. Transgenic animals thus generated may be tested with potential therapeutic or toxic agents.

Polynucleotides encoding CADECM may also be manipulated *in vitro* in ES cells derived from human blastocysts. Human ES cells have the potential to differentiate into at least eight separate cell lineages including endoderm, mesoderm, and ectodermal cell types. These cell lineages differentiate into, for example, neural cells, hematopoietic lineages, and cardiomyocytes (Thomson, J.A. et al. (1998) Science 282:1145-1147).

Polynucleotides encoding CADECM can also be used to create "knockin" humanized animals (pigs) or transgenic animals (mice or rats) to model human disease. With knockin technology, a region of a polynucleotide encoding CADECM is injected into animal ES cells, and the injected sequence integrates into the animal cell genome. Transformed cells are injected into blastulae, and the blastulae are implanted as described above. Transgenic progeny or inbred lines are studied and treated with potential pharmaceutical agents to obtain information on treatment of a human disease. Alternatively, a mammal inbred to overexpress CADECM, e.g., by secreting CADECM in its milk, may also serve as a convenient source of that protein (Janne, J. et al. (1998) Biotechnol. Annu. Rev. 4:55-74).

THERAPEUTICS

Chemical and structural similarity, e.g., in the context of sequences and motifs, exists between regions of CADECM and cell adhesion and extracellular matrix proteins. In addition, examples of tissues expressing CADECM can be found in Table 6 and can also be found in Example XI. Therefore, CADECM appears to play a role in immune system disorders, neurological disorders, developmental disorders, connective tissue disorders, and cell proliferative disorders, including cancer. In the treatment of disorders associated with increased CADECM expression or activity, it is desirable to decrease the expression or activity of CADECM. In the treatment of disorders associated with decreased CADECM expression or activity, it is desirable to increase the expression or activity of CADECM.

Therefore, in one embodiment, CADECM or a fragment or derivative thereof may be administered to a subject to treat or prevent a disorder associated with decreased expression or activity of CADECM. Examples of such disorders include, but are not limited to, an immune system disorder, such as acquired immunodeficiency syndrome (AIDS), X-linked agammaglobinemia of Bruton, common variable immunodeficiency (CVI), DiGeorge's syndrome (thymic hypoplasia), thymic dysplasia, isolated IgA deficiency, severe combined immunodeficiency disease (SCID), immunodeficiency with thrombocytopenia and eczema (Wiskott-Aldrich syndrome), Chediak-Higashi syndrome, chronic granulomatous diseases, hereditary angioneurotic edema, immunodeficiency associated with Cushing's disease, Addison's disease, adult respiratory distress syndrome, allergies, ankylosing spondylitis, amyloidosis, anemia, asthma, atherosclerosis, autoimmune hemolytic anemia,

autoimmune thyroiditis, autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy (APECED), bronchitis, cholecystitis, contact dermatitis, Crohn's disease, atopic dermatitis, dermatomyositis, diabetes mellitus, emphysema, episodic lymphopenia with lymphocytotoxins, erythroblastosis fetalis, erythema nodosum, atrophic gastritis, glomerulonephritis, Goodpasture's syndrome, gout, Graves' disease, Hashimoto's thyroiditis, hypereosinophilia, irritable bowel syndrome, multiple sclerosis, myasthenia gravis, myocardial or pericardial inflammation, osteoarthritis, osteoporosis, pancreatitis, polymyositis, psoriasis, Reiter's syndrome, rheumatoid arthritis, scleroderma, Sjögren's syndrome, systemic anaphylaxis, systemic lupus erythematosus, systemic sclerosis, thrombocytopenic purpura, ulcerative colitis, uveitis, Werner syndrome, complications of cancer, hemodialysis, and extracorporeal circulation, viral, bacterial, fungal, parasitic, protozoal, and helminthic infections, and trauma; a neurological disorder, such as epilepsy, ischemic cerebrovascular disease, stroke, cerebral neoplasms, Alzheimer's disease, Pick's disease, Huntington's disease, dementia, Parkinson's disease and other extrapyramidal disorders, amyotrophic lateral sclerosis and other motor neuron disorders, progressive neural muscular atrophy, retinitis pigmentosa, hereditary ataxias, multiple sclerosis and other demyelinating diseases, bacterial and viral meningitis, brain abscess, subdural empyema, epidural abscess, suppurative intracranial thrombophlebitis, myelitis and radiculitis, viral central nervous system disease, prion diseases including kuru, Creutzfeldt-Jakob disease, and Gerstmann-Straussler-Scheinker syndrome, fatal familial insomnia, nutritional and metabolic diseases of the nervous system, neurofibromatosis, tuberous sclerosis, cerebelloretinal hemangioblastomatosis, encephalotrigeminal syndrome, mental retardation and other developmental disorders of the central nervous system including Down syndrome, cerebral palsy, neuroskeletal disorders, autonomic nervous system disorders, cranial nerve disorders, spinal cord diseases, muscular dystrophy and other neuromuscular disorders, peripheral nervous system disorders, dermatomyositis and polymyositis, inherited, metabolic, endocrine, and toxic myopathies, myasthenia gravis, periodic paralysis, mental disorders including mood, anxiety, and schizophrenic disorders, seasonal affective disorder (SAD), akathesia, amnesia, catatonia, diabetic neuropathy, tardive dyskinesia, dystonias, paranoid psychoses, postherpetic neuralgia, Tourette's disorder, progressive supranuclear palsy, corticobasal degeneration, and familial frontotemporal dementia; a developmental disorder, such as renal tubular acidosis, anemia, Cushing's syndrome, achondroplastic dwarfism, Duchenne and Becker muscular dystrophy, epilepsy, gonadal dysgenesis, WAGR syndrome (Wilms' tumor, aniridia, genitourinary abnormalities, and mental retardation), Smith-Magenis syndrome, myelodysplastic syndrome, hereditary mucoepithelial dysplasia, hereditary keratodermas, hereditary neuropathies such as Charcot-Marie-Tooth disease and neurofibromatosis, hypothyroidism, hydrocephalus, seizure disorders such as Syndenham's chorea and

cerebral palsy, spina bifida, anencephaly, craniorachischisis, congenital glaucoma, cataract, and sensorineural hearing loss; a connective tissue disorder, such as osteogenesis imperfecta, Ehlers-Danlos syndrome, chondrodysplasias, Marfan syndrome, Alport syndrome, familial aortic aneurysm, achondroplasia, mucopolysaccharidoses, osteoporosis, osteopetrosis, Paget's disease, rickets, osteomalacia, hyperparathyroidism, renal osteodystrophy, osteonecrosis, osteomyelitis, osteoma, osteoid osteoma, osteoblastoma, osteosarcoma, osteochondroma, chondroma, chondroblastoma, chondromyxoid fibroma, chondrosarcoma, fibrous cortical defect, nonossifying fibroma, fibrous dysplasia, fibrosarcoma, malignant fibrous histiocytoma, Ewing's sarcoma, primitive neuroectodermal tumor, giant cell tumor, osteoarthritis, rheumatoid arthritis, ankylosing spondyloarthritis, Reiter's syndrome, psoriatic arthritis, enteropathic arthritis, infectious arthritis, gout, gouty arthritis, calcium pyrophosphate crystal deposition disease, ganglion, synovial cyst, villonodular synovitis, systemic sclerosis, Dupuytren's contracture, hepatic fibrosis, lupus erythematosus, mixed connective tissue disease, epidermolysis bullosa simplex, bullous congenital ichthyosiform erythroderma (epidermolytic hyperkeratosis), non-epidermolytic and epidermolytic palmoplantar keratoderma, ichthyosis bullosa of Siemens, pachyonychia congenita, and white sponge nevus; and a cell proliferative disorder, such as actinic keratosis, arteriosclerosis, atherosclerosis, bursitis, cirrhosis, hepatitis, mixed connective tissue disease (MCTD), myelofibrosis, paroxysmal nocturnal hemoglobinuria, polycythemia vera, psoriasis, primary thrombocythemia, Tangier disease, and cancers including adenocarcinoma, leukemia, lymphoma, melanoma, myeloma, sarcoma, teratocarcinoma, and, in particular, cancers of the adrenal gland, bladder, bone, bone marrow, brain, breast, cervix, colon, gall bladder, ganglia, gastrointestinal tract, heart, kidney, liver, lung, muscle, ovary, pancreas, parathyroid, penis, prostate, salivary glands, skin, spleen, testis, thymus, thyroid, and uterus.

In another embodiment, a vector capable of expressing CADECM or a fragment or derivative thereof may be administered to a subject to treat or prevent a disorder associated with decreased expression or activity of CADECM including, but not limited to, those described above.

In a further embodiment, a composition comprising a substantially purified CADECM in conjunction with a suitable pharmaceutical carrier may be administered to a subject to treat or prevent a disorder associated with decreased expression or activity of CADECM including, but not limited to, those provided above.

In still another embodiment, an agonist which modulates the activity of CADECM may be administered to a subject to treat or prevent a disorder associated with decreased expression or activity of CADECM including, but not limited to, those listed above.

In a further embodiment, an antagonist of CADECM may be administered to a subject to treat

or prevent a disorder associated with increased expression or activity of CADECM. Examples of such disorders include, but are not limited to, those immune system disorders, neurological disorders, developmental disorders, connective tissue disorders, and cell proliferative disorders, including cancer described above. In one aspect, an antibody which specifically binds CADECM may be used directly as an antagonist or indirectly as a targeting or delivery mechanism for bringing a pharmaceutical agent to cells or tissues which express CADECM.

In an additional embodiment, a vector expressing the complement of the polynucleotide encoding CADECM may be administered to a subject to treat or prevent a disorder associated with increased expression or activity of CADECM including, but not limited to, those described above.

In other embodiments, any protein, agonist, antagonist, antibody, complementary sequence, or vector embodiments may be administered in combination with other appropriate therapeutic agents. Selection of the appropriate agents for use in combination therapy may be made by one of ordinary skill in the art, according to conventional pharmaceutical principles. The combination of therapeutic agents may act synergistically to effect the treatment or prevention of the various disorders described above. Using this approach, one may be able to achieve therapeutic efficacy with lower dosages of each agent, thus reducing the potential for adverse side effects.

An antagonist of CADECM may be produced using methods which are generally known in the art. In particular, purified CADECM may be used to produce antibodies or to screen libraries of pharmaceutical agents to identify those which specifically bind CADECM. Antibodies to CADECM may also be generated using methods that are well known in the art. Such antibodies may include, but are not limited to, polyclonal, monoclonal, chimeric, and single chain antibodies, Fab fragments, and fragments produced by a Fab expression library. In an embodiment, neutralizing antibodies (i.e., those which inhibit dimer formation) can be used therapeutically. Single chain antibodies (e.g., from camels or llamas) may be potent enzyme inhibitors and may have application in the design of peptide mimetics, and in the development of immuno-adsorbents and biosensors (Muyldermans, S. (2001) J. Biotechnol. 74:277-302).

For the production of antibodies, various hosts including goats, rabbits, rats, mice, camels, dromedaries, llamas, humans, and others may be immunized by injection with CADECM or with any fragment or oligopeptide thereof which has immunogenic properties. Depending on the host species, various adjuvants may be used to increase immunological response. Such adjuvants include, but are not limited to, Freund's, mineral gels such as aluminum hydroxide, and surface active substances such as lysolecithin, pluronic polyols, polyanions, peptides, oil emulsions, KLH, and dinitrophenol. Among adjuvants used in humans, BCG (bacilli Calmette-Guerin) and *Corynebacterium parvum* are

especially preferable.

It is preferred that the oligopeptides, peptides, or fragments used to induce antibodies to CADECM have an amino acid sequence consisting of at least about 5 amino acids, and generally will consist of at least about 10 amino acids. It is also preferable that these oligopeptides, peptides, or fragments are substantially identical to a portion of the amino acid sequence of the natural protein. Short stretches of CADECM amino acids may be fused with those of another protein, such as KLH, and antibodies to the chimeric molecule may be produced.

Monoclonal antibodies to CADECM may be prepared using any technique which provides for the production of antibody molecules by continuous cell lines in culture. These include, but are not limited to, the hybridoma technique, the human B-cell hybridoma technique, and the EBV-hybridoma technique (Kohler, G. et al. (1975) Nature 256:495-497; Kozbor, D. et al. (1985) J. Immunol. Methods 81:31-42; Cote, R.J. et al. (1983) Proc. Natl. Acad. Sci. USA 80:2026-2030; Cole, S.P. et al. (1984) Mol. Cell Biol. 62:109-120).

In addition, techniques developed for the production of "chimeric antibodies," such as the splicing of mouse antibody genes to human antibody genes to obtain a molecule with appropriate antigen specificity and biological activity, can be used (Morrison, S.L. et al. (1984) Proc. Natl. Acad. Sci. USA 81:6851-6855; Neuberger, M.S. et al. (1984) Nature 312:604-608; Takeda, S. et al. (1985) Nature 314:452-454). Alternatively, techniques described for the production of single chain antibodies may be adapted, using methods known in the art, to produce CADECM-specific single chain antibodies. Antibodies with related specificity, but of distinct idiotypic composition, may be generated by chain shuffling from random combinatorial immunoglobulin libraries (Burton, D.R. (1991) Proc. Natl. Acad. Sci. USA 88:10134-10137).

Antibodies may also be produced by inducing *in vivo* production in the lymphocyte population or by screening immunoglobulin libraries or panels of highly specific binding reagents as disclosed in the literature (Orlandi, R. et al. (1989) Proc. Natl. Acad. Sci. USA 86:3833-3837; Winter, G. et al. (1991) Nature 349:293-299).

Antibody fragments which contain specific binding sites for CADECM may also be generated. For example, such fragments include, but are not limited to, $F(ab')_2$ fragments produced by pepsin digestion of the antibody molecule and Fab fragments generated by reducing the disulfide bridges of the F(ab')2 fragments. Alternatively, Fab expression libraries may be constructed to allow rapid and easy identification of monoclonal Fab fragments with the desired specificity (Huse, W.D. et al. (1989) Science 246:1275-1281).

Various immunoassays may be used for screening to identify antibodies having the desired

specificity. Numerous protocols for competitive binding or immunoradiometric assays using either polyclonal or monoclonal antibodies with established specificities are well known in the art. Such immunoassays typically involve the measurement of complex formation between CADECM and its specific antibody. A two-site, monoclonal-based immunoassay utilizing monoclonal antibodies reactive to two non-interfering CADECM epitopes is generally used, but a competitive binding assay may also be employed (Pound, *supra*).

Various methods such as Scatchard analysis in conjunction with radioimmunoassay techniques may be used to assess the affinity of antibodies for CADECM. Affinity is expressed as an association constant, K_a, which is defined as the molar concentration of CADECM-antibody complex divided by the molar concentrations of free antigen and free antibody under equilibrium conditions. The K_a determined for a preparation of polyclonal antibodies, which are heterogeneous in their affinities for multiple CADECM epitopes, represents the average affinity, or avidity, of the antibodies for CADECM. The K_a determined for a preparation of monoclonal antibodies, which are monospecific for a particular CADECM epitope, represents a true measure of affinity. High-affinity antibody preparations with K_a ranging from about 10^9 to 10^{12} L/mole are preferred for use in immunoassays in which the CADECM-antibody complex must withstand rigorous manipulations. Low-affinity antibody preparations with K_a ranging from about 10^6 to 10^7 L/mole are preferred for use in immunopurification and similar procedures which ultimately require dissociation of CADECM, preferably in active form, from the antibody (Catty, D. (1988) Antibodies, Volume I: A Practical Approach, IRL Press, Washington DC; Liddell, J.E. and A. Cryer (1991) A Practical Guide to Monoclonal Antibodies, John Wiley & Sons, New York NY).

The titer and avidity of polyclonal antibody preparations may be further evaluated to determine the quality and suitability of such preparations for certain downstream applications. For example, a polyclonal antibody preparation containing at least 1-2 mg specific antibody/ml, preferably 5-10 mg specific antibody/ml, is generally employed in procedures requiring precipitation of CADECM-antibody complexes. Procedures for evaluating antibody specificity, titer, and avidity, and guidelines for antibody quality and usage in various applications, are generally available (Catty, *supra*; Coligan et al., *supra*).

In another embodiment of the invention, polynucleotides encoding CADECM, or any fragment or complement thereof, may be used for therapeutic purposes. In one aspect, modifications of gene expression can be achieved by designing complementary sequences or antisense molecules (DNA, RNA, PNA, or modified oligonucleotides) to the coding or regulatory regions of the gene encoding CADECM. Such technology is well known in the art, and antisense oligonucleotides or larger

fragments can be designed from various locations along the coding or control regions of sequences encoding CADECM (Agrawal, S., ed. (1996) Antisense Therapeutics, Humana Press, Totawa NJ).

In therapeutic use, any gene delivery system suitable for introduction of the antisense sequences into appropriate target cells can be used. Antisense sequences can be delivered intracellularly in the form of an expression plasmid which, upon transcription, produces a sequence complementary to at least a portion of the cellular sequence encoding the target protein (Slater, J.E. et al. (1998) J. Allergy Clin. Immunol. 102:469-475; Scanlon, K.J. et al. (1995) FASEB J. 9:1288-1296). Antisense sequences can also be introduced intracellularly through the use of viral vectors, such as retrovirus and adeno-associated virus vectors (Miller, A.D. (1990) Blood 76:271-278; Ausubel et al., *supra*; Uckert, W. and W. Walther (1994) Pharmacol. Ther. 63:323-347). Other gene delivery mechanisms include liposome-derived systems, artificial viral envelopes, and other systems known in the art (Rossi, J.J. (1995) Br. Med. Bull. 51:217-225; Boado, R.J. et al. (1998) J. Pharm. Sci. 87:1308-1315; Morris, M.C. et al. (1997) Nucleic Acids Res. 25:2730-2736).

In another embodiment of the invention, polynucleotides encoding CADECM may be used for somatic or germline gene therapy. Gene therapy may be performed to (i) correct a genetic deficiency (e.g., in the cases of severe combined immunodeficiency (SCID)-X1 disease characterized by X-linked inheritance (Cavazzana-Calvo, M. et al. (2000) Science 288:669-672), severe combined immunodeficiency syndrome associated with an inherited adenosine deaminase (ADA) deficiency (Blaese, R.M. et al. (1995) Science 270:475-480; Bordignon, C. et al. (1995) Science 270:470-475), cystic fibrosis (Zabner, J. et al. (1993) Cell 75:207-216; Crystal, R.G. et al. (1995) Hum. Gene Therapy 6:643-666; Crystal, R.G. et al. (1995) Hum. Gene Therapy 6:667-703), thalassamias, familial hypercholesterolemia, and hemophilia resulting from Factor VIII or Factor IX deficiencies (Crystal, R.G. (1995) Science 270:404-410; Verma, I.M. and N. Somia (1997) Nature 389:239-242)), (ii) express a conditionally lethal gene product (e.g., in the case of cancers which result from unregulated cell proliferation), or (iii) express a protein which affords protection against intracellular parasites (e.g., against human retroviruses, such as human immunodeficiency virus (HIV) (Baltimore, D. (1988) Nature 335:395-396; Poeschla, E. et al. (1996) Proc. Natl. Acad. Sci. USA 93:11395-11399), hepatitis B or C virus (HBV, HCV); fungal parasites, such as *Candida albicans* and *Paracoccidioides brasiliensis*; and protozoan parasites such as *Plasmodium falciparum* and *Trypanosoma cruzi*). In the case where a genetic deficiency in CADECM expression or regulation causes disease, the expression of CADECM from an appropriate population of transduced cells may alleviate the clinical manifestations caused by the genetic deficiency.

In a further embodiment of the invention, diseases or disorders caused by deficiencies in

CADECM are treated by constructing mammalian expression vectors encoding CADECM and introducing these vectors by mechanical means into CADECM-deficient cells. Mechanical transfer technologies for use with cells *in vivo* or *ex vitro* include (i) direct DNA microinjection into individual cells, (ii) ballistic gold particle delivery, (iii) liposome-mediated transfection, (iv) receptor-mediated gene transfer, and (v) the use of DNA transposons (Morgan, R.A. and W.F. Anderson (1993) Annu. Rev. Biochem. 62:191-217; Ivics, Z. (1997) Cell 91:501-510; Boulay, J.-L. and H. Récipon (1998) Curr. Opin. Biotechnol. 9:445-450).

Expression vectors that may be effective for the expression of CADECM include, but are not limited to, the PCDNA 3.1, EPITAG, PRCCMV2, PREP, PVAX, PCR2-TOPOTA vectors (Invitrogen, Carlsbad CA), PCMV-SCRIPT, PCMV-TAG, PEGSH/PERV (Stratagene, La Jolla CA), and PTET-OFF, PTET-ON, PTRE2, PTRE2-LUC, PTK-HYG (BD Clontech, Palo Alto CA). CADECM may be expressed using (i) a constitutively active promoter, (e.g., from cytomegalovirus (CMV), Rous sarcoma virus (RSV), SV40 virus, thymidine kinase (TK), or β-actin genes), (ii) an inducible promoter (e.g., the tetracycline-regulated promoter (Gossen, M. and H. Bujard (1992) Proc. Natl. Acad. Sci. USA 89:5547-5551; Gossen, M. et al. (1995) Science 268:1766-1769; Rossi, F.M.V. and H.M. Blau (1998) Curr. Opin. Biotechnol. 9:451-456), commercially available in the T-REX plasmid (Invitrogen)); the ecdysone-inducible promoter (available in the plasmids PVGRXR and PIND; Invitrogen); the FK506/rapamycin inducible promoter; or the RU486/mifepristone inducible promoter (Rossi, F.M.V. and H.M. Blau, *supra*)), or (iii) a tissue-specific promoter or the native promoter of the endogenous gene encoding CADECM from a normal individual.

Commercially available liposome transformation kits (e.g., the PERFECT LIPID TRANSFECTION KIT, available from Invitrogen) allow one with ordinary skill in the art to deliver polynucleotides to target cells in culture and require minimal effort to optimize experimental parameters. In the alternative, transformation is performed using the calcium phosphate method (Graham, F.L. and A.J. Eb (1973) Virology 52:456-467), or by electroporation (Neumann, E. et al. (1982) EMBO J. 1:841-845). The introduction of DNA to primary cells requires modification of these standardized mammalian transfection protocols.

In another embodiment of the invention, diseases or disorders caused by genetic defects with respect to CADECM expression are treated by constructing a retrovirus vector consisting of (i) the polynucleotide encoding CADECM under the control of an independent promoter or the retrovirus long terminal repeat (LTR) promoter, (ii) appropriate RNA packaging signals, and (iii) a Rev-responsive element (RRE) along with additional retrovirus *cis*-acting RNA sequences and coding sequences required for efficient vector propagation. Retrovirus vectors (e.g., PFB and PFBNEO) are

commercially available (Stratagene) and are based on published data (Riviere, I. et al. (1995) Proc. Natl. Acad. Sci. USA 92:6733-6737), incorporated by reference herein. The vector is propagated in an appropriate vector producing cell line (VPCL) that expresses an envelope gene with a tropism for receptors on the target cells or a promiscuous envelope protein such as VSVg (Armentano, D. et al. (1987) J. Virol. 61:1647-1650; Bender, M.A. et al. (1987) J. Virol. 61:1639-1646; Adam, M.A. and A.D. Miller (1988) J. Virol. 62:3802-3806; Dull, T. et al. (1998) J. Virol. 72:8463-8471; Zufferey, R. et al. (1998) J. Virol. 72:9873-9880). U.S. Patent No. 5,910,434 to Rigg ("Method for obtaining retrovirus packaging cell lines producing high transducing efficiency retroviral supernatant") discloses a method for obtaining retrovirus packaging cell lines and is hereby incorporated by reference. Propagation of retrovirus vectors, transduction of a population of cells (e.g., $CD4^+$ T-cells), and the return of transduced cells to a patient are procedures well known to persons skilled in the art of gene therapy and have been well documented (Ranga, U. et al. (1997) J. Virol. 71:7020-7029; Bauer, G. et al. (1997) Blood 89:2259-2267; Bonyhadi, M.L. (1997) J. Virol. 71:4707-4716; Ranga, U. et al. (1998) Proc. Natl. Acad. Sci. USA 95:1201-1206; Su, L. (1997) Blood 89:2283-2290).

In an embodiment, an adenovirus-based gene therapy delivery system is used to deliver polynucleotides encoding CADECM to cells which have one or more genetic abnormalities with respect to the expression of CADECM. The construction and packaging of adenovirus-based vectors are well known to those with ordinary skill in the art. Replication defective adenovirus vectors have proven to be versatile for importing genes encoding immunoregulatory proteins into intact islets in the pancreas (Csete, M.E. et al. (1995) Transplantation 27:263-268). Potentially useful adenoviral vectors are described in U.S. Patent No. 5,707,618 to Armentano ("Adenovirus vectors for gene therapy"), hereby incorporated by reference. For adenoviral vectors, see also Antinozzi, P.A. et al. (1999; Annu. Rev. Nutr. 19:511-544) and Verma, I.M. and N. Somia (1997; Nature 18:389:239-242).

In another embodiment, a herpes-based, gene therapy delivery system is used to deliver polynucleotides encoding CADECM to target cells which have one or more genetic abnormalities with respect to the expression of CADECM. The use of herpes simplex virus (HSV)-based vectors may be especially valuable for introducing CADECM to cells of the central nervous system, for which HSV has a tropism. The construction and packaging of herpes-based vectors are well known to those with ordinary skill in the art. A replication-competent herpes simplex virus (HSV) type 1-based vector has been used to deliver a reporter gene to the eyes of primates (Liu, X. et al. (1999) Exp. Eye Res. 169:385-395). The construction of a HSV-1 virus vector has also been disclosed in detail in U.S. Patent No. 5,804,413 to DeLuca ("Herpes simplex virus strains for gene transfer"), which is hereby incorporated by reference. U.S. Patent No. 5,804,413 teaches the use of recombinant HSV d92

which consists of a genome containing at least one exogenous gene to be transferred to a cell under the control of the appropriate promoter for purposes including human gene therapy. Also taught by this patent are the construction and use of recombinant HSV strains deleted for ICP4, ICP27 and ICP22. For HSV vectors, see also Goins, W.F. et al. (1999; J. Virol. 73:519-532) and Xu, H. et al. (1994; Dev. Biol. 163:152-161). The manipulation of cloned herpesvirus sequences, the generation of recombinant virus following the transfection of multiple plasmids containing different segments of the large herpesvirus genomes, the growth and propagation of herpesvirus, and the infection of cells with herpesvirus are techniques well known to those of ordinary skill in the art.

In another embodiment, an alphavirus (positive, single-stranded RNA virus) vector is used to deliver polynucleotides encoding CADECM to target cells. The biology of the prototypic alphavirus, Semliki Forest Virus (SFV), has been studied extensively and gene transfer vectors have been based on the SFV genome (Garoff, H. and K.-J. Li (1998) Curr. Opin. Biotechnol. 9:464-469). During alphavirus RNA replication, a subgenomic RNA is generated that normally encodes the viral capsid proteins. This subgenomic RNA replicates to higher levels than the full length genomic RNA, resulting in the overproduction of capsid proteins relative to the viral proteins with enzymatic activity (e.g., protease and polymerase). Similarly, inserting the coding sequence for CADECM into the alphavirus genome in place of the capsid-coding region results in the production of a large number of CADECM-coding RNAs and the synthesis of high levels of CADECM in vector transduced cells. While alphavirus infection is typically associated with cell lysis within a few days, the ability to establish a persistent infection in hamster normal kidney cells (BHK-21) with a variant of Sindbis virus (SIN) indicates that the lytic replication of alphaviruses can be altered to suit the needs of the gene therapy application (Dryga, S.A. et al. (1997) Virology 228:74-83). The wide host range of alphaviruses will allow the introduction of CADECM into a variety of cell types. The specific transduction of a subset of cells in a population may require the sorting of cells prior to transduction. The methods of manipulating infectious cDNA clones of alphaviruses, performing alphavirus cDNA and RNA transfections, and performing alphavirus infections, are well known to those with ordinary skill in the art.

Oligonucleotides derived from the transcription initiation site, e.g., between about positions -10 and +10 from the start site, may also be employed to inhibit gene expression. Similarly, inhibition can be achieved using triple helix base-pairing methodology. Triple helix pairing is useful because it causes inhibition of the ability of the double helix to open sufficiently for the binding of polymerases, transcription factors, or regulatory molecules. Recent therapeutic advances using triplex DNA have been described in the literature (Gee, J.E. et al. (1994) in Huber, B.E. and B.I. Carr, Molecular and

Immunologic Approaches, Futura Publishing, Mt. Kisco NY, pp. 163-177). A complementary sequence or antisense molecule may also be designed to block translation of mRNA by preventing the transcript from binding to ribosomes.

Ribozymes, enzymatic RNA molecules, may also be used to catalyze the specific cleavage of RNA. The mechanism of ribozyme action involves sequence-specific hybridization of the ribozyme molecule to complementary target RNA, followed by endonucleolytic cleavage. For example, engineered hammerhead motif ribozyme molecules may specifically and efficiently catalyze endonucleolytic cleavage of RNA molecules encoding CADECM.

Specific ribozyme cleavage sites within any potential RNA target are initially identified by scanning the target molecule for ribozyme cleavage sites, including the following sequences: GUA, GUU, and GUC. Once identified, short RNA sequences of between 15 and 20 ribonucleotides, corresponding to the region of the target gene containing the cleavage site, may be evaluated for secondary structural features which may render the oligonucleotide inoperable. The suitability of candidate targets may also be evaluated by testing accessibility to hybridization with complementary oligonucleotides using ribonuclease protection assays.

Complementary ribonucleic acid molecules and ribozymes may be prepared by any method known in the art for the synthesis of nucleic acid molecules. These include techniques for chemically synthesizing oligonucleotides such as solid phase phosphoramidite chemical synthesis. Alternatively, RNA molecules may be generated by *in vitro* and *in vivo* transcription of DNA molecules encoding CADECM. Such DNA sequences may be incorporated into a wide variety of vectors with suitable RNA polymerase promoters such as T7 or SP6. Alternatively, these cDNA constructs that synthesize complementary RNA, constitutively or inducibly, can be introduced into cell lines, cells, or tissues.

RNA molecules may be modified to increase intracellular stability and half-life. Possible modifications include, but are not limited to, the addition of flanking sequences at the 5' and/or 3' ends of the molecule, or the use of phosphorothioate or 2' O-methyl rather than phosphodiesterase linkages within the backbone of the molecule. This concept is inherent in the production of PNAs and can be extended in all of these molecules by the inclusion of nontraditional bases such as inosine, queosine, and wybutosine, as well as acetyl-, methyl-, thio-, and similarly modified forms of adenine, cytidine, guanine, thymine, and uridine which are not as easily recognized by endogenous endonucleases.

In other embodiments of the invention, the expression of one or more selected polynucleotides of the present invention can be altered, inhibited, decreased, or silenced using RNA interference (RNAi) or post-transcriptional gene silencing (PTGS) methods known in the art. RNAi is a post-

transcriptional mode of gene silencing in which double-stranded RNA (dsRNA) introduced into a targeted cell specifically suppresses the expression of the homologous gene (i.e., the gene bearing the sequence complementary to the dsRNA). This effectively knocks out or substantially reduces the expression of the targeted gene. PTGS can also be accomplished by use of DNA or DNA fragments as well. RNAi methods are described by Fire, A. et al. (1998; Nature 391:806-811) and Gura, T. (2000; Nature 404:804-808). PTGS can also be initiated by introduction of a complementary segment of DNA into the selected tissue using gene delivery and/or viral vector delivery methods described herein or known in the art.

RNAi can be induced in mammalian cells by the use of small interfering RNA also known as siRNA. siRNA are shorter segments of dsRNA (typically about 21 to 23 nucleotides in length) that result *in vivo* from cleavage of introduced dsRNA by the action of an endogenous ribonuclease. siRNA appear to be the mediators of the RNAi effect in mammals. The most effective siRNAs appear to be 21 nucleotide dsRNAs with 2 nucleotide 3' overhangs. The use of siRNA for inducing RNAi in mammalian cells is described by Elbashir, S.M. et al. (2001; Nature 411:494-498).

siRNA can be generated indirectly by introduction of dsRNA into the targeted cell. Alternatively, siRNA can be synthesized directly and introduced into a cell by transfection methods and agents described herein or known in the art (such as liposome-mediated transfection, viral vector methods, or other polynucleotide delivery/introductory methods). Suitable siRNAs can be selected by examining a transcript of the target polynucleotide (e.g., mRNA) for nucleotide sequences downstream from the AUG start codon and recording the occurrence of each nucleotide and the 3' adjacent 19 to 23 nucleotides as potential siRNA target sites, with sequences having a 21 nucleotide length being preferred. Regions to be avoided for target siRNA sites include the 5' and 3' untranslated regions (UTRs) and regions near the start codon (within 75 bases), as these may be richer in regulatory protein binding sites. UTR-binding proteins and/or translation initiation complexes may interfere with binding of the siRNP endonuclease complex. The selected target sites for siRNA can then be compared to the appropriate genome database (e.g., human, etc.) using BLAST or other sequence comparison algorithms known in the art. Target sequences with significant homology to other coding sequences can be eliminated from consideration. The selected siRNAs can be produced by chemical synthesis methods known in the art or by *in vitro* transcription using commercially available methods and kits such as the SILENCER siRNA construction kit (Ambion, Austin TX).

In alternative embodiments, long-term gene silencing and/or RNAi effects can be induced in selected tissue using expression vectors that continuously express siRNA. This can be accomplished using expression vectors that are engineered to express hairpin RNAs (shRNAs) using methods

known in the art (see, e.g., Brummelkamp, T.R. et al. (2002) Science 296:550-553; and Paddison, P.J. et al. (2002) Genes Dev. 16:948-958). In these and related embodiments, shRNAs can be delivered to target cells using expression vectors known in the art. An example of a suitable expression vector for delivery of siRNA is the PSILENCER1.0-U6 (circular) plasmid (Ambion). Once delivered to the target tissue, shRNAs are processed *in vivo* into siRNA-like molecules capable of carrying out gene-specific silencing.

In various embodiments, the expression levels of genes targeted by RNAi or PTGS methods can be determined by assays for mRNA and/or protein analysis. Expression levels of the mRNA of a targeted gene can be determined, for example, by northern analysis methods using the NORTHERNMAX-GLY kit (Ambion); by microarray methods; by PCR methods; by real time PCR methods; and by other RNA/polynucleotide assays known in the art or described herein. Expression levels of the protein encoded by the targeted gene can be determined, for example, by microarray methods; by polyacrylamide gel electrophoresis; and by Western analysis using standard techniques known in the art.

An additional embodiment of the invention encompasses a method for screening for a compound which is effective in altering expression of a polynucleotide encoding CADECM. Compounds which may be effective in altering expression of a specific polynucleotide may include, but are not limited to, oligonucleotides, antisense oligonucleotides, triple helix-forming oligonucleotides, transcription factors and other polypeptide transcriptional regulators, and non-macromolecular chemical entities which are capable of interacting with specific polynucleotide sequences. Effective compounds may alter polynucleotide expression by acting as either inhibitors or promoters of polynucleotide expression. Thus, in the treatment of disorders associated with increased CADECM expression or activity, a compound which specifically inhibits expression of the polynucleotide encoding CADECM may be therapeutically useful, and in the treatment of disorders associated with decreased CADECM expression or activity, a compound which specifically promotes expression of the polynucleotide encoding CADECM may be therapeutically useful.

In various embodiments, one or more test compounds may be screened for effectiveness in altering expression of a specific polynucleotide. A test compound may be obtained by any method commonly known in the art, including chemical modification of a compound known to be effective in altering polynucleotide expression; selection from an existing, commercially-available or proprietary library of naturally-occurring or non-natural chemical compounds; rational design of a compound based on chemical and/or structural properties of the target polynucleotide; and selection from a library of chemical compounds created combinatorially or randomly. A sample comprising a

polynucleotide encoding CADECM is exposed to at least one test compound thus obtained. The sample may comprise, for example, an intact or permeabilized cell, or an *in vitro* cell-free or reconstituted biochemical system. Alterations in the expression of a polynucleotide encoding CADECM are assayed by any method commonly known in the art. Typically, the expression of a specific nucleotide is detected by hybridization with a probe having a nucleotide sequence complementary to the sequence of the polynucleotide encoding CADECM. The amount of hybridization may be quantified, thus forming the basis for a comparison of the expression of the polynucleotide both with and without exposure to one or more test compounds. Detection of a change in the expression of a polynucleotide exposed to a test compound indicates that the test compound is effective in altering the expression of the polynucleotide. A screen for a compound effective in altering expression of a specific polynucleotide can be carried out, for example, using a *Schizosaccharomyces pombe* gene expression system (Atkins, D. et al. (1999) U.S. Patent No. 5,932,435; Arndt, G.M. et al. (2000) Nucleic Acids Res. 28:E15) or a human cell line such as HeLa cell (Clarke, M.L. et al. (2000) Biochem. Biophys. Res. Commun. 268:8-13). A particular embodiment of the present invention involves screening a combinatorial library of oligonucleotides (such as deoxyribonucleotides, ribonucleotides, peptide nucleic acids, and modified oligonucleotides) for antisense activity against a specific polynucleotide sequence (Bruice, T.W. et al. (1997) U.S. Patent No. 5,686,242; Bruice, T.W. et al. (2000) U.S. Patent No. 6,022,691).

Many methods for introducing vectors into cells or tissues are available and equally suitable for use *in vivo*, *in vitro*, and *ex vivo*. For *ex vivo* therapy, vectors may be introduced into stem cells taken from the patient and clonally propagated for autologous transplant back into that same patient. Delivery by transfection, by liposome injections, or by polycationic amino polymers may be achieved using methods which are well known in the art (Goldman, C.K. et al. (1997) Nat. Biotechnol. 15:462-466).

Any of the therapeutic methods described above may be applied to any subject in need of such therapy, including, for example, mammals such as humans, dogs, cats, cows, horses, rabbits, and monkeys.

An additional embodiment of the invention relates to the administration of a composition which generally comprises an active ingredient formulated with a pharmaceutically acceptable excipient. Excipients may include, for example, sugars, starches, celluloses, gums, and proteins. Various formulations are commonly known and are thoroughly discussed in the latest edition of Remington's Pharmaceutical Sciences (Maack Publishing, Easton PA). Such compositions may consist of CADECM, antibodies to CADECM, and mimetics, agonists, antagonists, or inhibitors of CADECM.

In various embodiments, the compositions described herein, such as pharmaceutical compositions, may be administered by any number of routes including, but not limited to, oral, intravenous, intramuscular, intra-arterial, intramedullary, intrathecal, intraventricular, pulmonary, transdermal, subcutaneous, intraperitoneal, intranasal, enteral, topical, sublingual, or rectal means.

Compositions for pulmonary administration may be prepared in liquid or dry powder form. These compositions are generally aerosolized immediately prior to inhalation by the patient. In the case of small molecules (e.g. traditional low molecular weight organic drugs), aerosol delivery of fast-acting formulations is well-known in the art. In the case of macromolecules (e.g. larger peptides and proteins), recent developments in the field of pulmonary delivery via the alveolar region of the lung have enabled the practical delivery of drugs such as insulin to blood circulation (see, e.g., Patton, J.S. et al., U.S. Patent No. 5,997,848). Pulmonary delivery allows administration without needle injection, and obviates the need for potentially toxic penetration enhancers.

Compositions suitable for use in the invention include compositions wherein the active ingredients are contained in an effective amount to achieve the intended purpose. The determination of an effective dose is well within the capability of those skilled in the art.

Specialized forms of compositions may be prepared for direct intracellular delivery of macromolecules comprising CADECM or fragments thereof. For example, liposome preparations containing a cell-impermeable macromolecule may promote cell fusion and intracellular delivery of the macromolecule. Alternatively, CADECM or a fragment thereof may be joined to a short cationic N-terminal portion from the HIV Tat-1 protein. Fusion proteins thus generated have been found to transduce into the cells of all tissues, including the brain, in a mouse model system (Schwarze, S.R. et al. (1999) Science 285:1569-1572).

For any compound, the therapeutically effective dose can be estimated initially either in cell culture assays, e.g., of neoplastic cells, or in animal models such as mice, rats, rabbits, dogs, monkeys, or pigs. An animal model may also be used to determine the appropriate concentration range and route of administration. Such information can then be used to determine useful doses and routes for administration in humans.

A therapeutically effective dose refers to that amount of active ingredient, for example CADECM or fragments thereof, antibodies of CADECM, and agonists, antagonists or inhibitors of CADECM, which ameliorates the symptoms or condition. Therapeutic efficacy and toxicity may be determined by standard pharmaceutical procedures in cell cultures or with experimental animals, such as by calculating the ED_{50} (the dose therapeutically effective in 50% of the population) or LD_{50} (the dose lethal to 50% of the population) statistics. The dose ratio of toxic to therapeutic effects is the

therapeutic index, which can be expressed as the LD_{50}/ED_{50} ratio. Compositions which exhibit large therapeutic indices are preferred. The data obtained from cell culture assays and animal studies are used to formulate a range of dosage for human use. The dosage contained in such compositions is preferably within a range of circulating concentrations that includes the ED_{50} with little or no toxicity. The dosage varies within this range depending upon the dosage form employed, the sensitivity of the patient, and the route of administration.

The exact dosage will be determined by the practitioner, in light of factors related to the subject requiring treatment. Dosage and administration are adjusted to provide sufficient levels of the active moiety or to maintain the desired effect. Factors which may be taken into account include the severity of the disease state, the general health of the subject, the age, weight, and gender of the subject, time and frequency of administration, drug combination(s), reaction sensitivities, and response to therapy. Long-acting compositions may be administered every 3 to 4 days, every week, or biweekly depending on the half-life and clearance rate of the particular formulation.

Normal dosage amounts may vary from about 0.1 μg to 100,000 μg, up to a total dose of about 1 gram, depending upon the route of administration. Guidance as to particular dosages and methods of delivery is provided in the literature and generally available to practitioners in the art. Those skilled in the art will employ different formulations for nucleotides than for proteins or their inhibitors. Similarly, delivery of polynucleotides or polypeptides will be specific to particular cells, conditions, locations, etc.

DIAGNOSTICS

In another embodiment, antibodies which specifically bind CADECM may be used for the diagnosis of disorders characterized by expression of CADECM, or in assays to monitor patients being treated with CADECM or agonists, antagonists, or inhibitors of CADECM. Antibodies useful for diagnostic purposes may be prepared in the same manner as described above for therapeutics. Diagnostic assays for CADECM include methods which utilize the antibody and a label to detect CADECM in human body fluids or in extracts of cells or tissues. The antibodies may be used with or without modification, and may be labeled by covalent or non-covalent attachment of a reporter molecule. A wide variety of reporter molecules, several of which are described above, are known in the art and may be used.

A variety of protocols for measuring CADECM, including ELISAs, RIAs, and FACS, are known in the art and provide a basis for diagnosing altered or abnormal levels of CADECM expression. Normal or standard values for CADECM expression are established by combining body fluids or cell extracts taken from normal mammalian subjects, for example, human subjects, with

antibodies to CADECM under conditions suitable for complex formation. The amount of standard complex formation may be quantitated by various methods, such as photometric means. Quantities of CADECM expressed in subject, control, and disease samples from biopsied tissues are compared with the standard values. Deviation between standard and subject values establishes the parameters for diagnosing disease.

In another embodiment of the invention, polynucleotides encoding CADECM may be used for diagnostic purposes. The polynucleotides which may be used include oligonucleotides, complementary RNA and DNA molecules, and PNAs. The polynucleotides may be used to detect and quantify gene expression in biopsied tissues in which expression of CADECM may be correlated with disease. The diagnostic assay may be used to determine absence, presence, and excess expression of CADECM, and to monitor regulation of CADECM levels during therapeutic intervention.

In one aspect, hybridization with PCR probes which are capable of detecting polynucleotides, including genomic sequences, encoding CADECM or closely related molecules may be used to identify nucleic acid sequences which encode CADECM. The specificity of the probe, whether it is made from a highly specific region, e.g., the 5' regulatory region, or from a less specific region, e.g., a conserved motif, and the stringency of the hybridization or amplification will determine whether the probe identifies only naturally occurring sequences encoding CADECM, allelic variants, or related sequences.

Probes may also be used for the detection of related sequences, and may have at least 50% sequence identity to any of the CADECM encoding sequences. The hybridization probes of the subject invention may be DNA or RNA and may be derived from the sequence of SEQ ID NO:43-84 or from genomic sequences including promoters, enhancers, and introns of the CADECM gene.

Means for producing specific hybridization probes for polynucleotides encoding CADECM include the cloning of polynucleotides encoding CADECM or CADECM derivatives into vectors for the production of mRNA probes. Such vectors are known in the art, are commercially available, and may be used to synthesize RNA probes *in vitro* by means of the addition of the appropriate RNA polymerases and the appropriate labeled nucleotides. Hybridization probes may be labeled by a variety of reporter groups, for example, by radionuclides such as ^{32}P or ^{35}S, or by enzymatic labels, such as alkaline phosphatase coupled to the probe via avidin/biotin coupling systems, and the like.

Polynucleotides encoding CADECM may be used for the diagnosis of disorders associated with expression of CADECM. Examples of such disorders include, but are not limited to, an immune system disorder, such as acquired immunodeficiency syndrome (AIDS), X-linked agammaglobinemia of Bruton, common variable immunodeficiency (CVI), DiGeorge's syndrome (thymic hypoplasia),

thymic dysplasia, isolated IgA deficiency, severe combined immunodeficiency disease (SCID), immunodeficiency with thrombocytopenia and eczema (Wiskott-Aldrich syndrome), Chediak-Higashi syndrome, chronic granulomatous diseases, hereditary angioneurotic edema, immunodeficiency associated with Cushing's disease, Addison's disease, adult respiratory distress syndrome, allergies, ankylosing spondylitis, amyloidosis, anemia, asthma, atherosclerosis, autoimmune hemolytic anemia, autoimmune thyroiditis, autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy (APECED), bronchitis, cholecystitis, contact dermatitis, Crohn's disease, atopic dermatitis, dermatomyositis, diabetes mellitus, emphysema, episodic lymphopenia with lymphocytotoxins, erythroblastosis fetalis, erythema nodosum, atrophic gastritis, glomerulonephritis, Goodpasture's syndrome, gout, Graves' disease, Hashimoto's thyroiditis, hypereosinophilia, irritable bowel syndrome, multiple sclerosis, myasthenia gravis, myocardial or pericardial inflammation, osteoarthritis, osteoporosis, pancreatitis, polymyositis, psoriasis, Reiter's syndrome, rheumatoid arthritis, scleroderma, Sjögren's syndrome, systemic anaphylaxis, systemic lupus erythematosus, systemic sclerosis, thrombocytopenic purpura, ulcerative colitis, uveitis, Werner syndrome, complications of cancer, hemodialysis, and extracorporeal circulation, viral, bacterial, fungal, parasitic, protozoal, and helminthic infections, and trauma; a neurological disorder, such as epilepsy, ischemic cerebrovascular disease, stroke, cerebral neoplasms, Alzheimer's disease, Pick's disease, Huntington's disease, dementia, Parkinson's disease and other extrapyramidal disorders, amyotrophic lateral sclerosis and other motor neuron disorders, progressive neural muscular atrophy, retinitis pigmentosa, hereditary ataxias, multiple sclerosis and other demyelinating diseases, bacterial and viral meningitis, brain abscess, subdural empyema, epidural abscess, suppurative intracranial thrombophlebitis, myelitis and radiculitis, viral central nervous system disease, prion diseases including kuru, Creutzfeldt-Jakob disease, and Gerstmann-Straussler-Scheinker syndrome, fatal familial insomnia, nutritional and metabolic diseases of the nervous system, neurofibromatosis, tuberous sclerosis, cerebelloretinal hemangioblastomatosis, encephalotrigeminal syndrome, mental retardation and other developmental disorders of the central nervous system including Down syndrome, cerebral palsy, neuroskeletal disorders, autonomic nervous system disorders, cranial nerve disorders, spinal cord diseases, muscular dystrophy and other neuromuscular disorders, peripheral nervous system disorders, dermatomyositis and polymyositis, inherited, metabolic, endocrine, and toxic myopathies, myasthenia gravis, periodic paralysis, mental disorders including mood, anxiety, and schizophrenic disorders, seasonal affective disorder (SAD), akathesia, amnesia, catatonia, diabetic neuropathy, tardive dyskinesia, dystonias, paranoid psychoses, postherpetic neuralgia, Tourette's disorder, progressive supranuclear palsy, corticobasal degeneration, and familial frontotemporal dementia; a developmental disorder, such as renal tubular acidosis, anemia,

Cushing's syndrome, achondroplastic dwarfism, Duchenne and Becker muscular dystrophy, epilepsy, gonadal dysgenesis, WAGR syndrome (Wilms' tumor, aniridia, genitourinary abnormalities, and mental retardation), Smith-Magenis syndrome, myelodysplastic syndrome, hereditary mucoepithelial dysplasia, hereditary keratodermas, hereditary neuropathies such as Charcot-Marie-Tooth disease and neurofibromatosis, hypothyroidism, hydrocephalus, seizure disorders such as Syndenham's chorea and cerebral palsy, spina bifida, anencephaly, craniorachischisis, congenital glaucoma, cataract, and sensorineural hearing loss; a connective tissue disorder, such as osteogenesis imperfecta, Ehlers-Danlos syndrome, chondrodysplasias, Marfan syndrome, Alport syndrome, familial aortic aneurysm, achondroplasia, mucopolysaccharidoses, osteoporosis, osteopetrosis, Paget's disease, rickets, osteomalacia, hyperparathyroidism, renal osteodystrophy, osteonecrosis, osteomyelitis, osteoma, osteoid osteoma, osteoblastoma, osteosarcoma, osteochondroma, chondroma, chondroblastoma, chondromyxoid fibroma, chondrosarcoma, fibrous cortical defect, nonossifying fibroma, fibrous dysplasia, fibrosarcoma, malignant fibrous histiocytoma, Ewing's sarcoma, primitive neuroectodermal tumor, giant cell tumor, osteoarthritis, rheumatoid arthritis, ankylosing spondyloarthritis, Reiter's syndrome, psoriatic arthritis, enteropathic arthritis, infectious arthritis, gout, gouty arthritis, calcium pyrophosphate crystal deposition disease, ganglion, synovial cyst, villonodular synovitis, systemic sclerosis, Dupuytren's contracture, hepatic fibrosis, lupus erythematosus, mixed connective tissue disease, epidermolysis bullosa simplex, bullous congenital ichthyosiform erythroderma (epidermolytic hyperkeratosis), non-epidermolytic and epidermolytic palmoplantar keratoderma, ichthyosis bullosa of Siemens, pachyonychia congenita, and white sponge nevus; and a cell proliferative disorder, such as actinic keratosis, arteriosclerosis, atherosclerosis, bursitis, cirrhosis, hepatitis, mixed connective tissue disease (MCTD), myelofibrosis, paroxysmal nocturnal hemoglobinuria, polycythemia vera, psoriasis, primary thrombocythemia, Tangier disease, and cancers including adenocarcinoma, leukemia, lymphoma, melanoma, myeloma, sarcoma, teratocarcinoma, and, in particular, cancers of the adrenal gland, bladder, bone, bone marrow, brain, breast, cervix, colon, gall bladder, ganglia, gastrointestinal tract, heart, kidney, liver, lung, muscle, ovary, pancreas, parathyroid, penis, prostate, salivary glands, skin, spleen, testis, thymus, thyroid, and uterus. Polynucleotides encoding CADECM may be used in Southern or northern analysis, dot blot, or other membrane-based technologies; in PCR technologies; in dipstick, pin, and multiformat ELISA-like assays; and in microarrays utilizing fluids or tissues from patients to detect altered CADECM expression. Such qualitative or quantitative methods are well known in the art.

In a particular embodiment, polynucleotides encoding CADECM may be used in assays that detect the presence of associated disorders, particularly those mentioned above. Polynucleotides

complementary to sequences encoding CADECM may be labeled by standard methods and added to a fluid or tissue sample from a patient under conditions suitable for the formation of hybridization complexes. After a suitable incubation period, the sample is washed and the signal is quantified and compared with a standard value. If the amount of signal in the patient sample is significantly altered in comparison to a control sample then the presence of altered levels of polynucleotides encoding CADECM in the sample indicates the presence of the associated disorder. Such assays may also be used to evaluate the efficacy of a particular therapeutic treatment regimen in animal studies, in clinical trials, or to monitor the treatment of an individual patient.

In order to provide a basis for the diagnosis of a disorder associated with expression of CADECM, a normal or standard profile for expression is established. This may be accomplished by combining body fluids or cell extracts taken from normal subjects, either animal or human, with a sequence, or a fragment thereof, encoding CADECM, under conditions suitable for hybridization or amplification. Standard hybridization may be quantified by comparing the values obtained from normal subjects with values from an experiment in which a known amount of a substantially purified polynucleotide is used. Standard values obtained in this manner may be compared with values obtained from samples from patients who are symptomatic for a disorder. Deviation from standard values is used to establish the presence of a disorder.

Once the presence of a disorder is established and a treatment protocol is initiated, hybridization assays may be repeated on a regular basis to determine if the level of expression in the patient begins to approximate that which is observed in the normal subject. The results obtained from successive assays may be used to show the efficacy of treatment over a period ranging from several days to months.

With respect to cancer, the presence of an abnormal amount of transcript (either under- or overexpressed) in biopsied tissue from an individual may indicate a predisposition for the development of the disease, or may provide a means for detecting the disease prior to the appearance of actual clinical symptoms. A more definitive diagnosis of this type may allow health professionals to employ preventative measures or aggressive treatment earlier, thereby preventing the development or further progression of the cancer.

Additional diagnostic uses for oligonucleotides designed from the sequences encoding CADECM may involve the use of PCR. These oligomers may be chemically synthesized, generated enzymatically, or produced *in vitro*. Oligomers will preferably contain a fragment of a polynucleotide encoding CADECM, or a fragment of a polynucleotide complementary to the polynucleotide encoding CADECM, and will be employed under optimized conditions for identification of a specific gene or

condition. Oligomers may also be employed under less stringent conditions for detection or quantification of closely related DNA or RNA sequences.

In a particular aspect, oligonucleotide primers derived from polynucleotides encoding CADECM may be used to detect single nucleotide polymorphisms (SNPs). SNPs are substitutions, insertions and deletions that are a frequent cause of inherited or acquired genetic disease in humans. Methods of SNP detection include, but are not limited to, single-stranded conformation polymorphism (SSCP) and fluorescent SSCP (fSSCP) methods. In SSCP, oligonucleotide primers derived from polynucleotides encoding CADECM are used to amplify DNA using the polymerase chain reaction (PCR). The DNA may be derived, for example, from diseased or normal tissue, biopsy samples, bodily fluids, and the like. SNPs in the DNA cause differences in the secondary and tertiary structures of PCR products in single-stranded form, and these differences are detectable using gel electrophoresis in non-denaturing gels. In fSCCP, the oligonucleotide primers are fluorescently labeled, which allows detection of the amplimers in high-throughput equipment such as DNA sequencing machines. Additionally, sequence database analysis methods, termed in silico SNP (isSNP), are capable of identifying polymorphisms by comparing the sequence of individual overlapping DNA fragments which assemble into a common consensus sequence. These computer-based methods filter out sequence variations due to laboratory preparation of DNA and sequencing errors using statistical models and automated analyses of DNA sequence chromatograms. In the alternative, SNPs may be detected and characterized by mass spectrometry using, for example, the high throughput MASSARRAY system (Sequenom, Inc., San Diego CA).

SNPs may be used to study the genetic basis of human disease. For example, at least 16 common SNPs have been associated with non-insulin-dependent diabetes mellitus. SNPs are also useful for examining differences in disease outcomes in monogenic disorders, such as cystic fibrosis, sickle cell anemia, or chronic granulomatous disease. For example, variants in the mannose-binding lectin, MBL2, have been shown to be correlated with deleterious pulmonary outcomes in cystic fibrosis. SNPs also have utility in pharmacogenomics, the identification of genetic variants that influence a patient's response to a drug, such as life-threatening toxicity. For example, a variation in N-acetyl transferase is associated with a high incidence of peripheral neuropathy in response to the anti-tuberculosis drug isoniazid, while a variation in the core promoter of the ALOX5 gene results in diminished clinical response to treatment with an anti-asthma drug that targets the 5-lipoxygenase pathway. Analysis of the distribution of SNPs in different populations is useful for investigating genetic drift, mutation, recombination, and selection, as well as for tracing the origins of populations and their migrations (Taylor, J.G. et al. (2001) Trends Mol. Med. 7:507-512; Kwok, P.-Y. and Z. Gu

(1999) Mol. Med. Today 5:538-543; Nowotny, P. et al. (2001) Curr. Opin. Neurobiol. 11:637-641).

Methods which may also be used to quantify the expression of CADECM include radiolabeling or biotinylating nucleotides, coamplification of a control nucleic acid, and interpolating results from standard curves (Melby, P.C. et al. (1993) J. Immunol. Methods 159:235-244; Duplaa, C. et al. (1993) Anal. Biochem. 212:229-236). The speed of quantitation of multiple samples may be accelerated by running the assay in a high-throughput format where the oligomer or polynucleotide of interest is presented in various dilutions and a spectrophotometric or colorimetric response gives rapid quantitation.

In further embodiments, oligonucleotides or longer fragments derived from any of the polynucleotides described herein may be used as elements on a microarray. The microarray can be used in transcript imaging techniques which monitor the relative expression levels of large numbers of genes simultaneously as described below. The microarray may also be used to identify genetic variants, mutations, and polymorphisms. This information may be used to determine gene function, to understand the genetic basis of a disorder, to diagnose a disorder, to monitor progression/regression of disease as a function of gene expression, and to develop and monitor the activities of therapeutic agents in the treatment of disease. In particular, this information may be used to develop a pharmacogenomic profile of a patient in order to select the most appropriate and effective treatment regimen for that patient. For example, therapeutic agents which are highly effective and display the fewest side effects may be selected for a patient based on his/her pharmacogenomic profile.

In another embodiment, CADECM, fragments of CADECM, or antibodies specific for CADECM may be used as elements on a microarray. The microarray may be used to monitor or measure protein-protein interactions, drug-target interactions, and gene expression profiles, as described above.

A particular embodiment relates to the use of the polynucleotides of the present invention to generate a transcript image of a tissue or cell type. A transcript image represents the global pattern of gene expression by a particular tissue or cell type. Global gene expression patterns are analyzed by quantifying the number of expressed genes and their relative abundance under given conditions and at a given time (Seilhamer et al., "Comparative Gene Transcript Analysis," U.S. Patent No. 5,840,484; hereby expressly incorporated by reference herein). Thus a transcript image may be generated by hybridizing the polynucleotides of the present invention or their complements to the totality of transcripts or reverse transcripts of a particular tissue or cell type. In one embodiment, the hybridization takes place in high-throughput format, wherein the polynucleotides of the present invention or their complements comprise a subset of a plurality of elements on a microarray. The

resultant transcript image would provide a profile of gene activity.

Transcript images may be generated using transcripts isolated from tissues, cell lines, biopsies, or other biological samples. The transcript image may thus reflect gene expression *in vivo*, as in the case of a tissue or biopsy sample, or *in vitro*, as in the case of a cell line.

Transcript images which profile the expression of the polynucleotides of the present invention may also be used in conjunction with *in vitro* model systems and preclinical evaluation of pharmaceuticals, as well as toxicological testing of industrial and naturally-occurring environmental compounds. All compounds induce characteristic gene expression patterns, frequently termed molecular fingerprints or toxicant signatures, which are indicative of mechanisms of action and toxicity (Nuwaysir, E.F. et al. (1999) Mol. Carcinog. 24:153-159; Steiner, S. and N.L. Anderson (2000) Toxicol. Lett. 112-113:467-471). If a test compound has a signature similar to that of a compound with known toxicity, it is likely to share those toxic properties. These fingerprints or signatures are most useful and refined when they contain expression information from a large number of genes and gene families. Ideally, a genome-wide measurement of expression provides the highest quality signature. Even genes whose expression is not altered by any tested compounds are important as well, as the levels of expression of these genes are used to normalize the rest of the expression data. The normalization procedure is useful for comparison of expression data after treatment with different compounds. While the assignment of gene function to elements of a toxicant signature aids in interpretation of toxicity mechanisms, knowledge of gene function is not necessary for the statistical matching of signatures which leads to prediction of toxicity (see, for example, Press Release 00-02 from the National Institute of Environmental Health Sciences, released February 29, 2000, available at niehs.nih.gov/oc/news/toxchip.htm). Therefore, it is important and desirable in toxicological screening using toxicant signatures to include all expressed gene sequences.

In an embodiment, the toxicity of a test compound can be assessed by treating a biological sample containing nucleic acids with the test compound. Nucleic acids that are expressed in the treated biological sample are hybridized with one or more probes specific to the polynucleotides of the present invention, so that transcript levels corresponding to the polynucleotides of the present invention may be quantified. The transcript levels in the treated biological sample are compared with levels in an untreated biological sample. Differences in the transcript levels between the two samples are indicative of a toxic response caused by the test compound in the treated sample.

Another embodiment relates to the use of the polypeptides disclosed herein to analyze the proteome of a tissue or cell type. The term proteome refers to the global pattern of protein expression in a particular tissue or cell type. Each protein component of a proteome can be subjected individually

to further analysis. Proteome expression patterns, or profiles, are analyzed by quantifying the number of expressed proteins and their relative abundance under given conditions and at a given time. A profile of a cell's proteome may thus be generated by separating and analyzing the polypeptides of a particular tissue or cell type. In one embodiment, the separation is achieved using two-dimensional gel electrophoresis, in which proteins from a sample are separated by isoelectric focusing in the first dimension, and then according to molecular weight by sodium dodecyl sulfate slab gel electrophoresis in the second dimension (Steiner and Anderson, *supra*). The proteins are visualized in the gel as discrete and uniquely positioned spots, typically by staining the gel with an agent such as Coomassie Blue or silver or fluorescent stains. The optical density of each protein spot is generally proportional to the level of the protein in the sample. The optical densities of equivalently positioned protein spots from different samples, for example, from biological samples either treated or untreated with a test compound or therapeutic agent, are compared to identify any changes in protein spot density related to the treatment. The proteins in the spots are partially sequenced using, for example, standard methods employing chemical or enzymatic cleavage followed by mass spectrometry. The identity of the protein in a spot may be determined by comparing its partial sequence, preferably of at least 5 contiguous amino acid residues, to the polypeptide sequences of interest. In some cases, further sequence data may be obtained for definitive protein identification.

A proteomic profile may also be generated using antibodies specific for CADECM to quantify the levels of CADECM expression. In one embodiment, the antibodies are used as elements on a microarray, and protein expression levels are quantified by contacting the microarray with the sample and detecting the levels of protein bound to each array element (Lueking, A. et al. (1999) Anal. Biochem. 270:103-111; Mendoze, L.G. et al. (1999) Biotechniques 27:778-788). Detection may be performed by a variety of methods known in the art, for example, by reacting the proteins in the sample with a thiol- or amino-reactive fluorescent compound and detecting the amount of fluorescence bound at each array element.

Toxicant signatures at the proteome level are also useful for toxicological screening, and should be analyzed in parallel with toxicant signatures at the transcript level. There is a poor correlation between transcript and protein abundances for some proteins in some tissues (Anderson, N.L. and J. Seilhamer (1997) Electrophoresis 18:533-537), so proteome toxicant signatures may be useful in the analysis of compounds which do not significantly affect the transcript image, but which alter the proteomic profile. In addition, the analysis of transcripts in body fluids is difficult, due to rapid degradation of mRNA, so proteomic profiling may be more reliable and informative in such cases.

In another embodiment, the toxicity of a test compound is assessed by treating a biological

sample containing proteins with the test compound. Proteins that are expressed in the treated biological sample are separated so that the amount of each protein can be quantified. The amount of each protein is compared to the amount of the corresponding protein in an untreated biological sample. A difference in the amount of protein between the two samples is indicative of a toxic response to the test compound in the treated sample. Individual proteins are identified by sequencing the amino acid residues of the individual proteins and comparing these partial sequences to the polypeptides of the present invention.

In another embodiment, the toxicity of a test compound is assessed by treating a biological sample containing proteins with the test compound. Proteins from the biological sample are incubated with antibodies specific to the polypeptides of the present invention. The amount of protein recognized by the antibodies is quantified. The amount of protein in the treated biological sample is compared with the amount in an untreated biological sample. A difference in the amount of protein between the two samples is indicative of a toxic response to the test compound in the treated sample.

Microarrays may be prepared, used, and analyzed using methods known in the art (Brennan, T.M. et al. (1995) U.S. Patent No. 5,474,796; Schena, M. et al. (1996) Proc. Natl. Acad. Sci. USA 93:10614-10619; Baldeschweiler et al. (1995) PCT application WO95/25116; Shalon, D. et al. (1995) PCT application WO95/35505; Heller, R.A. et al. (1997) Proc. Natl. Acad. Sci. USA 94:2150-2155; Heller, M.J. et al. (1997) U.S. Patent No. 5,605,662). Various types of microarrays are well known and thoroughly described in Schena, M., ed. (1999; DNA Microarrays: A Practical Approach, Oxford University Press, London).

In another embodiment of the invention, nucleic acid sequences encoding CADECM may be used to generate hybridization probes useful in mapping the naturally occurring genomic sequence. Either coding or noncoding sequences may be used, and in some instances, noncoding sequences may be preferable over coding sequences. For example, conservation of a coding sequence among members of a multi-gene family may potentially cause undesired cross hybridization during chromosomal mapping. The sequences may be mapped to a particular chromosome, to a specific region of a chromosome, or to artificial chromosome constructions, e.g., human artificial chromosomes (HACs), yeast artificial chromosomes (YACs), bacterial artificial chromosomes (BACs), bacterial P1 constructions, or single chromosome cDNA libraries (Harrington, J.J. et al. (1997) Nat. Genet. 15:345-355; Price, C.M. (1993) Blood Rev. 7:127-134; Trask, B.J. (1991) Trends Genet. 7:149-154). Once mapped, the nucleic acid sequences may be used to develop genetic linkage maps, for example, which correlate the inheritance of a disease state with the inheritance of a particular chromosome region or restriction fragment length polymorphism (RFLP) (Lander, E.S. and D. Botstein (1986) Proc. Natl.

Acad. Sci. USA 83:7353-7357).

Fluorescent *in situ* hybridization (FISH) may be correlated with other physical and genetic map data (Heinz-Ulrich, et al. (1995) in Meyers, *supra*, pp. 965-968). Examples of genetic map data can be found in various scientific journals or at the Online Mendelian Inheritance in Man (OMIM) World Wide Web site. Correlation between the location of the gene encoding CADECM on a physical map and a specific disorder, or a predisposition to a specific disorder, may help define the region of DNA associated with that disorder and thus may further positional cloning efforts.

In situ hybridization of chromosomal preparations and physical mapping techniques, such as linkage analysis using established chromosomal markers, may be used for extending genetic maps. Often the placement of a gene on the chromosome of another mammalian species, such as mouse, may reveal associated markers even if the exact chromosomal locus is not known. This information is valuable to investigators searching for disease genes using positional cloning or other gene discovery techniques. Once the gene or genes responsible for a disease or syndrome have been crudely localized by genetic linkage to a particular genomic region, e.g., ataxia-telangiectasia to 11q22-23, any sequences mapping to that area may represent associated or regulatory genes for further investigation (Gatti, R.A. et al. (1988) Nature 336:577-580). The nucleotide sequence of the instant invention may also be used to detect differences in the chromosomal location due to translocation, inversion, etc., among normal, carrier, or affected individuals.

In another embodiment of the invention, CADECM, its catalytic or immunogenic fragments, or oligopeptides thereof can be used for screening libraries of compounds in any of a variety of drug screening techniques. The fragment employed in such screening may be free in solution, affixed to a solid support, borne on a cell surface, or located intracellularly. The formation of binding complexes between CADECM and the agent being tested may be measured.

Another technique for drug screening provides for high throughput screening of compounds having suitable binding affinity to the protein of interest (Geysen, et al. (1984) PCT application WO84/03564). In this method, large numbers of different small test compounds are synthesized on a solid substrate. The test compounds are reacted with CADECM, or fragments thereof, and washed. Bound CADECM is then detected by methods well known in the art. Purified CADECM can also be coated directly onto plates for use in the aforementioned drug screening techniques. Alternatively, non-neutralizing antibodies can be used to capture the peptide and immobilize it on a solid support.

In another embodiment, one may use competitive drug screening assays in which neutralizing antibodies capable of binding CADECM specifically compete with a test compound for binding CADECM. In this manner, antibodies can be used to detect the presence of any peptide which

shares one or more antigenic determinants with CADECM.

In additional embodiments, the nucleotide sequences which encode CADECM may be used in any molecular biology techniques that have yet to be developed, provided the new techniques rely on properties of nucleotide sequences that are currently known, including, but not limited to, such properties as the triplet genetic code and specific base pair interactions.

Without further elaboration, it is believed that one skilled in the art can, using the preceding description, utilize the present invention to its fullest extent. The following embodiments are, therefore. to be construed as merely illustrative, and not limitative of the remainder of the disclosure in any way whatsoever.

The disclosures of all patents, applications, and publications mentioned above and below, including U.S. Ser. No. 60/403,781, U.S. Ser. No. 60/407,034, U.S. Ser. No. 60/410,566, U.S. Ser. No. 60/413,482, U.S. Ser. No. 60/413,890, U.S. Ser. No. 60/424,904, and U.S. Ser. No. 60/426,222, are hereby expressly incorporated by reference.

EXAMPLES

I. Construction of cDNA Libraries

Incyte cDNAs are derived from cDNA libraries described in the LIFESEQ database (Incyte, Palo Alto CA). Some tissues are homogenized and lysed in guanidinium isothiocyanate, while others are homogenized and lysed in phenol or in a suitable mixture of denaturants, such as TRIZOL (Invitrogen), a monophasic solution of phenol and guanidine isothiocyanate. The resulting lysates are centrifuged over CsCl cushions or extracted with chloroform. RNA is precipitated from the lysates with either isopropanol or sodium acetate and ethanol, or by other routine methods.

Phenol extraction and precipitation of RNA are repeated as necessary to increase RNA purity. In some cases, RNA is treated with DNase. For most libraries, poly(A)+ RNA is isolated using oligo d(T)-coupled paramagnetic particles (Promega), OLIGOTEX latex particles (QIAGEN, Chatsworth CA), or an OLIGOTEX mRNA purification kit (QIAGEN). Alternatively, RNA is isolated directly from tissue lysates using other RNA isolation kits, e.g., the POLY(A)PURE mRNA purification kit (Ambion, Austin TX).

In some cases, Stratagene is provided with RNA and constructs the corresponding cDNA libraries. Otherwise, cDNA is synthesized and cDNA libraries are constructed with the UNIZAP vector system (Stratagene) or SUPERSCRIPT plasmid system (Invitrogen), using the recommended procedures or similar methods known in the art (Ausubel et al., *supra*, ch. 5). Reverse transcription is initiated using oligo d(T) or random primers. Synthetic oligonucleotide adapters are ligated to double

stranded cDNA, and the cDNA is digested with the appropriate restriction enzyme or enzymes. For most libraries, the cDNA is size-selected (300-1000 bp) using SEPHACRYL S1000, SEPHAROSE CL2B, or SEPHAROSE CL4B column chromatography (Amersham Biosciences) or preparative agarose gel electrophoresis. cDNAs are ligated into compatible restriction enzyme sites of the polylinker of a suitable plasmid, e.g., PBLUESCRIPT plasmid (Stratagene), PSPORT1 plasmid (Invitrogen, Carlsbad CA), PCDNA2.1 plasmid (Invitrogen), PBK-CMV plasmid (Stratagene), PCR2-TOPOTA plasmid (Invitrogen), PCMV-ICIS plasmid (Stratagene), pIGEN (Incyte, Palo Alto CA), pRARE (Incyte), or pINCY (Incyte), or derivatives thereof. Recombinant plasmids are transformed into competent *E. coli* cells including XL1-Blue, XL1-BlueMRF, or SOLR from Stratagene or DH5α, DH10B, or ElectroMAX DH10B from Invitrogen.

II. Isolation of cDNA Clones

Plasmids obtained as described in Example I are recovered from host cells by *in vivo* excision using the UNIZAP vector system (Stratagene) or by cell lysis. Plasmids are purified using at least one of the following: a Magic or WIZARD Minipreps DNA purification system (Promega); an AGTC Miniprep purification kit (Edge Biosystems, Gaithersburg MD); and QIAWELL 8 Plasmid, QIAWELL 8 Plus Plasmid, QIAWELL 8 Ultra Plasmid purification systems or the R.E.A.L. PREP 96 plasmid purification kit from QIAGEN. Following precipitation, plasmids are resuspended in 0.1 ml of distilled water and stored, with or without lyophilization, at 4°C.

Alternatively, plasmid DNA is amplified from host cell lysates using direct link PCR in a high-throughput format (Rao, V.B. (1994) Anal. Biochem. 216:1-14). Host cell lysis and thermal cycling steps are carried out in a single reaction mixture. Samples are processed and stored in 384-well plates, and the concentration of amplified plasmid DNA is quantified fluorometrically using PICOGREEN dye (Molecular Probes, Eugene OR) and a FLUOROSKAN II fluorescence scanner (Labsystems Oy, Helsinki, Finland).

III. Sequencing and Analysis

Incyte cDNA recovered in plasmids as described in Example II are sequenced as follows. Sequencing reactions are processed using standard methods or high-throughput instrumentation such as the ABI CATALYST 800 (Applied Biosystems) thermal cycler or the PTC-200 thermal cycler (MJ Research) in conjunction with the HYDRA microdispenser (Robbins Scientific) or the MICROLAB 2200 (Hamilton) liquid transfer system. cDNA sequencing reactions are prepared using reagents provided by Amersham Biosciences or supplied in ABI sequencing kits such as the ABI PRISM BIGDYE Terminator cycle sequencing ready reaction kit (Applied Biosystems). Electrophoretic separation of cDNA sequencing reactions and detection of labeled polynucleotides are

carried out using the MEGABACE 1000 DNA sequencing system (Amersham Biosciences); the ABI PRISM 373 or 377 sequencing system (Applied Biosystems) in conjunction with standard ABI protocols and base calling software; or other sequence analysis systems known in the art. Reading frames within the cDNA sequences are identified using standard methods (Ausubel et al., *supra*, ch. 7). Some of the cDNA sequences are selected for extension using the techniques disclosed in Example VIII.

Polynucleotide sequences derived from Incyte cDNAs are validated by removing vector, linker, and poly(A) sequences and by masking ambiguous bases, using algorithms and programs based on BLAST, dynamic programming, and dinucleotide nearest neighbor analysis. The Incyte cDNA sequences or translations thereof are then queried against a selection of public databases such as the GenBank primate, rodent, mammalian, vertebrate, and eukaryote databases, and BLOCKS, PRINTS, DOMO, PRODOM; PROTEOME databases with sequences from *Homo sapiens*, *Rattus norvegicus*, *Mus musculus*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and *Candida albicans* (Incyte, Palo Alto CA); hidden Markov model (HMM)-based protein family databases such as PFAM, INCY, and TIGRFAM (Haft, D.H. et al. (2001) Nucleic Acids Res. 29:41-43); and HMM-based protein domain databases such as SMART (Schultz, J. et al. (1998) Proc. Natl. Acad. Sci. USA 95:5857-5864; Letunic, I. et al. (2002) Nucleic Acids Res. 30:242-244). (HMM is a probabilistic approach which analyzes consensus primary structures of gene families; see, for example, Eddy, S.R. (1996) Curr. Opin. Struct. Biol. 6:361-365.) The queries are performed using programs based on BLAST, FASTA, BLIMPS, and HMMER. The Incyte cDNA sequences are assembled to produce full length polynucleotide sequences. Alternatively, GenBank cDNAs, GenBank ESTs, stitched sequences, stretched sequences, or Genscan-predicted coding sequences (see Examples IV and V) are used to extend Incyte cDNA assemblages to full length. Assembly is performed using programs based on Phred, Phrap, and Consed, and cDNA assemblages are screened for open reading frames using programs based on GeneMark, BLAST, and FASTA. The full length polynucleotide sequences are translated to derive the corresponding full length polypeptide sequences. Alternatively, a polypeptide may begin at any of the methionine residues of the full length translated polypeptide. Full length polypeptide sequences are subsequently analyzed by querying against databases such as the GenBank protein databases (genpept), SwissProt, the PROTEOME databases, BLOCKS, PRINTS, DOMO, PRODOM, Prosite, hidden Markov model (HMM)-based protein family databases such as PFAM, INCY, and TIGRFAM; and HMM-based protein domain databases such as SMART. Full length polynucleotide sequences are also analyzed using MACDNASIS PRO software (MiraiBio, Alameda CA) and

LASERGENE software (DNASTAR). Polynucleotide and polypeptide sequence alignments are generated using default parameters specified by the CLUSTAL algorithm as incorporated into the MEGALIGN multisequence alignment program (DNASTAR), which also calculates the percent identity between aligned sequences.

Table 7 summarizes tools, programs, and algorithms used for the analysis and assembly of Incyte cDNA and full length sequences and provides applicable descriptions, references, and threshold parameters. The first column of Table 7 shows the tools, programs, and algorithms used, the second column provides brief descriptions thereof, the third column presents appropriate references, all of which are incorporated by reference herein in their entirety, and the fourth column presents, where applicable, the scores, probability values, and other parameters used to evaluate the strength of a match between two sequences (the higher the score or the lower the probability value, the greater the identity between two sequences).

The programs described above for the assembly and analysis of full length polynucleotide and polypeptide sequences are also used to identify polynucleotide sequence fragments from SEQ ID NO:43-84. Fragments from about 20 to about 4000 nucleotides which are useful in hybridization and amplification technologies are described in Table 4, column 2.

IV. Identification and Editing of Coding Sequences from Genomic DNA

Putative cell adhesion and extracellular matrix proteins are initially identified by running the Genscan gene identification program against public genomic sequence databases (e.g., gbpri and gbhtg). Genscan is a general-purpose gene identification program which analyzes genomic DNA sequences from a variety of organisms (Burge, C. and S. Karlin (1997) J. Mol. Biol. 268:78-94; Burge, C. and S. Karlin (1998) Curr. Opin. Struct. Biol. 8:346-354). The program concatenates predicted exons to form an assembled cDNA sequence extending from a methionine to a stop codon. The output of Genscan is a FASTA database of polynucleotide and polypeptide sequences. The maximum range of sequence for Genscan to analyze at once is set to 30 kb. To determine which of these Genscan predicted cDNA sequences encode cell adhesion and extracellular matrix proteins, the encoded polypeptides are analyzed by querying against PFAM models for cell adhesion and extracellular matrix proteins. Potential cell adhesion and extracellular matrix proteins are also identified by homology to Incyte cDNA sequences that have been annotated as cell adhesion and extracellular matrix proteins. These selected Genscan-predicted sequences are then compared by BLAST analysis to the genpept and gbpri public databases. Where necessary, the Genscan-predicted sequences are then edited by comparison to the top BLAST hit from genpept to correct errors in the sequence predicted by Genscan, such as extra or omitted exons. BLAST analysis is also used to find

any Incyte cDNA or public cDNA coverage of the Genscan-predicted sequences, thus providing evidence for transcription. When Incyte cDNA coverage is available, this information is used to correct or confirm the Genscan predicted sequence. Full length polynucleotide sequences are obtained by assembling Genscan-predicted coding sequences with Incyte cDNA sequences and/or public cDNA sequences using the assembly process described in Example III. Alternatively, full length polynucleotide sequences are derived entirely from edited or unedited Genscan-predicted coding sequences.

V. Assembly of Genomic Sequence Data with cDNA Sequence Data

__"Stitched" Sequences__

Partial cDNA sequences are extended with exons predicted by the Genscan gene identification program described in Example IV. Partial cDNAs assembled as described in Example III are mapped to genomic DNA and parsed into clusters containing related cDNAs and Genscan exon predictions from one or more genomic sequences. Each cluster is analyzed using an algorithm based on graph theory and dynamic programming to integrate cDNA and genomic information, generating possible splice variants that are subsequently confirmed, edited, or extended to create a full length sequence. Sequence intervals in which the entire length of the interval is present on more than one sequence in the cluster are identified, and intervals thus identified are considered to be equivalent by transitivity. For example, if an interval is present on a cDNA and two genomic sequences, then all three intervals are considered to be equivalent. This process allows unrelated but consecutive genomic sequences to be brought together, bridged by cDNA sequence. Intervals thus identified are then "stitched" together by the stitching algorithm in the order that they appear along their parent sequences to generate the longest possible sequence, as well as sequence variants. Linkages between intervals which proceed along one type of parent sequence (cDNA to cDNA or genomic sequence to genomic sequence) are given preference over linkages which change parent type (cDNA to genomic sequence). The resultant stitched sequences are translated and compared by BLAST analysis to the genpept and gbpri public databases. Incorrect exons predicted by Genscan are corrected by comparison to the top BLAST hit from genpept. Sequences are further extended with additional cDNA sequences, or by inspection of genomic DNA, when necessary.

__"Stretched" Sequences__

Partial DNA sequences are extended to full length with an algorithm based on BLAST analysis. First, partial cDNAs assembled as described in Example III are queried against public databases such as the GenBank primate, rodent, mammalian, vertebrate, and eukaryote databases using the BLAST program. The nearest GenBank protein homolog is then compared by BLAST

analysis to either Incyte cDNA sequences or GenScan exon predicted sequences described in Example IV. A chimeric protein is generated by using the resultant high-scoring segment pairs (HSPs) to map the translated sequences onto the GenBank protein homolog. Insertions or deletions may occur in the chimeric protein with respect to the original GenBank protein homolog. The GenBank protein homolog, the chimeric protein, or both are used as probes to search for homologous genomic sequences from the public human genome databases. Partial DNA sequences are therefore "stretched" or extended by the addition of homologous genomic sequences. The resultant stretched sequences are examined to determine whether they contain a complete gene.

VI. Chromosomal Mapping of CADECM Encoding Polynucleotides

The sequences used to assemble SEQ ID NO:43-84 are compared with sequences from the Incyte LIFESEQ database and public domain databases using BLAST and other implementations of the Smith-Waterman algorithm. Sequences from these databases that matched SEQ ID NO:43-84 are assembled into clusters of contiguous and overlapping sequences using assembly algorithms such as Phrap (Table 7). Radiation hybrid and genetic mapping data available from public resources such as the Stanford Human Genome Center (SHGC), Whitehead Institute for Genome Research (WIGR), and Généthon are used to determine if any of the clustered sequences have been previously mapped. Inclusion of a mapped sequence in a cluster results in the assignment of all sequences of that cluster, including its particular SEQ ID NO:, to that map location.

Map locations are represented by ranges, or intervals, of human chromosomes. The map position of an interval, in centiMorgans, is measured relative to the terminus of the chromosome's p-arm. (The centiMorgan (cM) is a unit of measurement based on recombination frequencies between chromosomal markers. On average, 1 cM is roughly equivalent to 1 megabase (Mb) of DNA in humans, although this can vary widely due to hot and cold spots of recombination.) The cM distances are based on genetic markers mapped by Généthon which provide boundaries for radiation hybrid markers whose sequences were included in each of the clusters. Human genome maps and other resources available to the public, such as the NCBI "GeneMap'99" World Wide Web site (ncbi.nlm.nih.gov/genemap/), can be employed to determine if previously identified disease genes map within or in proximity to the intervals indicated above.

VII. Analysis of Polynucleotide Expression

Northern analysis is a laboratory technique used to detect the presence of a transcript of a gene and involves the hybridization of a labeled nucleotide sequence to a membrane on which RNAs from a particular cell type or tissue have been bound (Sambrook and Russell, *supra*, ch. 7; Ausubel et al., *supra*, ch. 4).

Analogous computer techniques applying BLAST are used to search for identical or related molecules in databases such as GenBank or LIFESEQ (Incyte). This analysis is much faster than multiple membrane-based hybridizations. In addition, the sensitivity of the computer search can be modified to determine whether any particular match is categorized as exact or similar. The basis of the search is the product score, which is defined as:

$$\frac{\text{BLAST Score x Percent Identity}}{5 \text{ x minimum } \{\text{length(Seq. 1), length(Seq. 2)}\}}$$

The product score takes into account both the degree of similarity between two sequences and the length of the sequence match. The product score is a normalized value between 0 and 100, and is calculated as follows: the BLAST score is multiplied by the percent nucleotide identity and the product is divided by (5 times the length of the shorter of the two sequences). The BLAST score is calculated by assigning a score of +5 for every base that matches in a high-scoring segment pair (HSP), and -4 for every mismatch. Two sequences may share more than one HSP (separated by gaps). If there is more than one HSP, then the pair with the highest BLAST score is used to calculate the product score. The product score represents a balance between fractional overlap and quality in a BLAST alignment. For example, a product score of 100 is produced only for 100% identity over the entire length of the shorter of the two sequences being compared. A product score of 70 is produced either by 100% identity and 70% overlap at one end, or by 88% identity and 100% overlap at the other. A product score of 50 is produced either by 100% identity and 50% overlap at one end, or 79% identity and 100% overlap.

Alternatively, polynucleotides encoding CADECM are analyzed with respect to the tissue sources from which they are derived. For example, some full length sequences are assembled, at least in part, with overlapping Incyte cDNA sequences (see Example III). Each cDNA sequence is derived from a cDNA library constructed from a human tissue. Each human tissue is classified into one of the following organ/tissue categories: cardiovascular system; connective tissue; digestive system; embryonic structures; endocrine system; exocrine glands; genitalia, female; genitalia, male; germ cells; hemic and immune system; liver; musculoskeletal system; nervous system; pancreas; respiratory system; sense organs; skin; stomatognathic system; unclassified/mixed; or urinary tract. The number of libraries in each category is counted and divided by the total number of libraries across all categories. Similarly, each human tissue is classified into one of the following disease/condition categories: cancer, cell line, developmental, inflammation, neurological, trauma, cardiovascular, pooled, and other, and the number of libraries in each category is counted and divided by the total number of

libraries across all categories. The resulting percentages reflect the tissue- and disease-specific expression of cDNA encoding CADECM. cDNA sequences and cDNA library/tissue information are found in the LIFESEQ database (Incyte, Palo Alto CA).

VIII. Extension of CADECM Encoding Polynucleotides

Full length polynucleotides are produced by extension of an appropriate fragment of the full length molecule using oligonucleotide primers designed from this fragment. One primer is synthesized to initiate 5' extension of the known fragment, and the other primer is synthesized to initiate 3' extension of the known fragment. The initial primers are designed using OLIGO 4.06 software (National Biosciences), or another appropriate program, to be about 22 to 30 nucleotides in length, to have a GC content of about 50% or more, and to anneal to the target sequence at temperatures of about 68°C to about 72°C. Any stretch of nucleotides which would result in hairpin structures and primer-primer dimerizations is avoided.

Selected human cDNA libraries are used to extend the sequence. If more than one extension is necessary or desired, additional or nested sets of primers are designed.

High fidelity amplification is obtained by PCR using methods well known in the art. PCR is performed in 96-well plates using the PTC-200 thermal cycler (MJ Research, Inc.). The reaction mix contains DNA template, 200 nmol of each primer, reaction buffer containing Mg^{2+}, $(NH_4)_2SO_4$, and 2-mercaptoethanol, Taq DNA polymerase (Amersham Biosciences), ELONGASE enzyme (Invitrogen), and Pfu DNA polymerase (Stratagene), with the following parameters for primer pair PCI A and PCI B: Step 1: 94°C, 3 min; Step 2: 94°C, 15 sec; Step 3: 60°C, 1 min; Step 4: 68°C, 2 min; Step 5: Steps 2, 3, and 4 repeated 20 times; Step 6: 68°C, 5 min; Step 7: storage at 4°C. In the alternative, the parameters for primer pair T7 and SK+ are as follows: Step 1: 94°C, 3 min; Step 2: 94°C, 15 sec; Step 3: 57°C, 1 min; Step 4: 68°C, 2 min; Step 5: Steps 2, 3, and 4 repeated 20 times; Step 6: 68°C, 5 min; Step 7: storage at 4°C.

The concentration of DNA in each well is determined by dispensing 100 μl PICOGREEN quantitation reagent (0.25% (v/v) PICOGREEN; Molecular Probes, Eugene OR) dissolved in 1X TE and 0.5 μl of undiluted PCR product into each well of an opaque fluorimeter plate (Corning Costar, Acton MA), allowing the DNA to bind to the reagent. The plate is scanned in a Fluoroskan II (Labsystems Oy, Helsinki, Finland) to measure the fluorescence of the sample and to quantify the concentration of DNA. A 5 μl to 10 μl aliquot of the reaction mixture is analyzed by electrophoresis on a 1 % agarose gel to determine which reactions are successful in extending the sequence.

The extended nucleotides are desalted and concentrated, transferred to 384-well plates, digested with CviJI cholera virus endonuclease (Molecular Biology Research, Madison WI), and

sonicated or sheared prior to religation into pUC 18 vector (Amersham Biosciences). For shotgun sequencing, the digested nucleotides are separated on low concentration (0.6 to 0.8%) agarose gels, fragments are excised, and agar digested with Agar ACE (Promega). Extended clones were religated using T4 ligase (New England Biolabs, Beverly MA) into pUC 18 vector (Amersham Biosciences), treated with Pfu DNA polymerase (Stratagene) to fill-in restriction site overhangs, and transfected into competent *E. coli* cells. Transformed cells are selected on antibiotic-containing media, and individual colonies are picked and cultured overnight at 37°C in 384-well plates in LB/2x carb liquid media.

The cells are lysed, and DNA is amplified by PCR using Taq DNA polymerase (Amersham Biosciences) and Pfu DNA polymerase (Stratagene) with the following parameters: Step 1: 94°C, 3 min; Step 2: 94°C, 15 sec; Step 3: 60°C, 1 min; Step 4: 72°C, 2 min; Step 5: steps 2, 3, and 4 repeated 29 times; Step 6: 72°C, 5 min; Step 7: storage at 4°C. DNA is quantified by PICOGREEN reagent (Molecular Probes) as described above. Samples with low DNA recoveries are reamplified using the same conditions as described above. Samples are diluted with 20% dimethysulfoxide (1:2, v/v), and sequenced using DYENAMIC energy transfer sequencing primers and the DYENAMIC DIRECT kit (Amersham Biosciences) or the ABI PRISM BIGDYE Terminator cycle sequencing ready reaction kit (Applied Biosystems).

In like manner, full length polynucleotides are verified using the above procedure or are used to obtain 5' regulatory sequences using the above procedure along with oligonucleotides designed for such extension, and an appropriate genomic library.

IX. Identification of Single Nucleotide Polymorphisms in CADECM Encoding Polynucleotides

Common DNA sequence variants known as single nucleotide polymorphisms (SNPs) are identified in SEQ ID NO:43-84 using the LIFESEQ database (Incyte). Sequences from the same gene are clustered together and assembled as described in Example III, allowing the identification of all sequence variants in the gene. An algorithm consisting of a series of filters is used to distinguish SNPs from other sequence variants. Preliminary filters remove the majority of basecall errors by requiring a minimum Phred quality score of 15, and remove sequence alignment errors and errors resulting from improper trimming of vector sequences, chimeras, and splice variants. An automated procedure of advanced chromosome analysis is applied to the original chromatogram files in the vicinity of the putative SNP. Clone error filters use statistically generated algorithms to identify errors introduced during laboratory processing, such as those caused by reverse transcriptase, polymerase, or somatic mutation. Clustering error filters use statistically generated algorithms to identify errors

resulting from clustering of close homologs or pseudogenes, or due to contamination by non-human sequences. A final set of filters removes duplicates and SNPs found in immunoglobulins or T-cell receptors.

Certain SNPs are selected for further characterization by mass spectrometry using the high throughput MASSARRAY system (Sequenom, Inc.) to analyze allele frequencies at the SNP sites in four different human populations. The Caucasian population comprises 92 individuals (46 male, 46 female), including 83 from Utah, four French, three Venezualan, and two Amish individuals. The African population comprises 194 individuals (97 male, 97 female), all African Americans. The Hispanic population comprises 324 individuals (162 male, 162 female), all Mexican Hispanic. The Asian population comprises 126 individuals (64 male, 62 female) with a reported parental breakdown of 43% Chinese, 31% Japanese, 13% Korean, 5% Vietnamese, and 8% other Asian. Allele frequencies are first analyzed in the Caucasian population; in some cases those SNPs which show no allelic variance in this population are not further tested in the other three populations.

X. Labeling and Use of Individual Hybridization Probes

Hybridization probes derived from SEQ ID NO:43-84 are employed to screen cDNAs, genomic DNAs, or mRNAs. Although the labeling of oligonucleotides, consisting of about 20 base pairs, is specifically described, essentially the same procedure is used with larger nucleotide fragments. Oligonucleotides are designed using state-of-the-art software such as OLIGO 4.06 software (National Biosciences) and labeled by combining 50 pmol of each oligomer, 250 μCi of [γ-^{32}P] adenosine triphosphate (Amersham Biosciences), and T4 polynucleotide kinase (DuPont NEN, Boston MA). The labeled oligonucleotides are substantially purified using a SEPHADEX G-25 superfine size exclusion dextran bead column (Amersham Biosciences). An aliquot containing 10^7 counts per minute of the labeled probe is used in a typical membrane-based hybridization analysis of human genomic DNA digested with one of the following endonucleases: Ase I, Bgl II, Eco RI, Pst I, Xba I, or Pvu II (DuPont NEN).

The DNA from each digest is fractionated on a 0.7% agarose gel and transferred to nylon membranes (Nytran Plus, Schleicher & Schuell, Durham NH). Hybridization is carried out for 16 hours at 40°C. To remove nonspecific signals, blots are sequentially washed at room temperature under conditions of up to, for example, 0.1 x saline sodium citrate and 0.5% sodium dodecyl sulfate. Hybridization patterns are visualized using autoradiography or an alternative imaging means and compared.

XI. Microarrays

The linkage or synthesis of array elements upon a microarray can be achieved utilizing

photolithography, piezoelectric printing (ink-jet printing; see, e.g., Baldeschweiler et al., *supra*), mechanical microspotting technologies, and derivatives thereof. The substrate in each of the aforementioned technologies should be uniform and solid with a non-porous surface (Schena, M., ed. (1999) DNA Microarrays: A Practical Approach, Oxford University Press, London). Suggested substrates include silicon, silica, glass slides, glass chips, and silicon wafers. Alternatively, a procedure analogous to a dot or slot blot may also be used to arrange and link elements to the surface of a substrate using thermal, UV, chemical, or mechanical bonding procedures. A typical array may be produced using available methods and machines well known to those of ordinary skill in the art and may contain any appropriate number of elements (Schena, M. et al. (1995) Science 270:467-470; Shalon, D. et al. (1996) Genome Res. 6:639-645; Marshall, A. and J. Hodgson (1998) Nat. Biotechnol. 16:27-31).

Full length cDNAs, Expressed Sequence Tags (ESTs), or fragments or oligomers thereof may comprise the elements of the microarray. Fragments or oligomers suitable for hybridization can be selected using software well known in the art such as LASERGENE software (DNASTAR). The array elements are hybridized with polynucleotides in a biological sample. The polynucleotides in the biological sample are conjugated to a fluorescent label or other molecular tag for ease of detection. After hybridization, nonhybridized nucleotides from the biological sample are removed, and a fluorescence scanner is used to detect hybridization at each array element. Alternatively, laser desorbtion and mass spectrometry may be used for detection of hybridization. The degree of complementarity and the relative abundance of each polynucleotide which hybridizes to an element on the microarray may be assessed. In one embodiment, microarray preparation and usage is described in detail below.

Tissue or Cell Sample Preparation

Total RNA is isolated from tissue samples using the guanidinium thiocyanate method and poly(A)$^+$ RNA is purified using the oligo-(dT) cellulose method. Each poly(A)$^+$ RNA sample is reverse transcribed using MMLV reverse-transcriptase, 0.05 pg/μl oligo-(dT) primer (21mer), 1X first strand buffer, 0.03 units/μl RNase inhibitor, 500 μM dATP, 500 μM dGTP, 500 μM dTTP, 40 μM dCTP, 40 μM dCTP-Cy3 (BDS) or dCTP-Cy5 (Amersham Biosciences). The reverse transcription reaction is performed in a 25 ml volume containing 200 ng poly(A)$^+$ RNA with GEMBRIGHT kits (Incyte). Specific control poly(A)$^+$ RNAs are synthesized by *in vitro* transcription from non-coding yeast genomic DNA. After incubation at 37° C for 2 hr, each reaction sample (one with Cy3 and another with Cy5 labeling) is treated with 2.5 ml of 0.5M sodium hydroxide and incubated for 20 minutes at 85° C to the stop the reaction and degrade the RNA. Samples are purified using two

successive CHROMA SPIN 30 gel filtration spin columns (BD Clontech, Palo Alto CA) and after combining, both reaction samples are ethanol precipitated using 1 ml of glycogen (1 mg/ml), 60 ml sodium acetate, and 300 ml of 100% ethanol. The sample is then dried to completion using a SpeedVAC (Savant Instruments Inc., Holbrook NY) and resuspended in 14 μl 5X SSC/0.2% SDS.

Microarray Preparation

Sequences of the present invention are used to generate array elements. Each array element is amplified from bacterial cells containing vectors with cloned cDNA inserts. PCR amplification uses primers complementary to the vector sequences flanking the cDNA insert. Array elements are amplified in thirty cycles of PCR from an initial quantity of 1-2 ng to a final quantity greater than 5 μg. Amplified array elements are then purified using SEPHACRYL-400 (Amersham Biosciences).

Purified array elements are immobilized on polymer-coated glass slides. Glass microscope slides (Corning) are cleaned by ultrasound in 0.1% SDS and acetone, with extensive distilled water washes between and after treatments. Glass slides are etched in 4% hydrofluoric acid (VWR Scientific Products Corporation (VWR), West Chester PA), washed extensively in distilled water, and coated with 0.05% aminopropyl silane (Sigma-Aldrich, St. Louis MO) in 95% ethanol. Coated slides are cured in a 110°C oven.

Array elements are applied to the coated glass substrate using a procedure described in U.S. Patent No. 5,807,522, incorporated herein by reference. 1 μl of the array element DNA, at an average concentration of 100 ng/μl, is loaded into the open capillary printing element by a high-speed robotic apparatus. The apparatus then deposits about 5 nl of array element sample per slide.

Microarrays are UV-crosslinked using a STRATALINKER UV-crosslinker (Stratagene). Microarrays are washed at room temperature once in 0.2% SDS and three times in distilled water. Non-specific binding sites are blocked by incubation of microarrays in 0.2% casein in phosphate buffered saline (PBS) (Tropix, Inc., Bedford MA) for 30 minutes at 60° C followed by washes in 0.2% SDS and distilled water as before.

Hybridization

Hybridization reactions contain 9 μl of sample mixture consisting of 0.2 μg each of Cy3 and Cy5 labeled cDNA synthesis products in 5X SSC, 0.2% SDS hybridization buffer. The sample mixture is heated to 65° C for 5 minutes and is aliquoted onto the microarray surface and covered with an 1.8 cm^2 coverslip. The arrays are transferred to a waterproof chamber having a cavity just slightly larger than a microscope slide. The chamber is kept at 100% humidity internally by the addition of 140 μl of 5X SSC in a corner of the chamber. The chamber containing the arrays is incubated for about 6.5 hours at 60° C. The arrays are washed for 10 min at 45° C in a first wash buffer (1X SSC, 0.1%

SDS), three times for 10 minutes each at 45° C in a second wash buffer (0.1X SSC), and dried.

Detection

Reporter-labeled hybridization complexes are detected with a microscope equipped with an Innova 70 mixed gas 10 W laser (Coherent, Inc., Santa Clara CA) capable of generating spectral lines at 488 nm for excitation of Cy3 and at 632 nm for excitation of Cy5. The excitation laser light is focused on the array using a 20X microscope objective (Nikon, Inc., Melville NY). The slide containing the array is placed on a computer-controlled X-Y stage on the microscope and raster-scanned past the objective. The 1.8 cm x 1.8 cm array used in the present example is scanned with a resolution of 20 micrometers.

In two separate scans, a mixed gas multiline laser excites the two fluorophores sequentially. Emitted light is split, based on wavelength, into two photomultiplier tube detectors (PMT R1477, Hamamatsu Photonics Systems, Bridgewater NJ) corresponding to the two fluorophores. Appropriate filters positioned between the array and the photomultiplier tubes are used to filter the signals. The emission maxima of the fluorophores used are 565 nm for Cy3 and 650 nm for Cy5. Each array is typically scanned twice, one scan per fluorophore using the appropriate filters at the laser source, although the apparatus is capable of recording the spectra from both fluorophores simultaneously.

The sensitivity of the scans is typically calibrated using the signal intensity generated by a cDNA control species added to the sample mixture at a known concentration. A specific location on the array contains a complementary DNA sequence, allowing the intensity of the signal at that location to be correlated with a weight ratio of hybridizing species of 1:100,000. When two samples from different sources (e.g., representing test and control cells), each labeled with a different fluorophore, are hybridized to a single array for the purpose of identifying genes that are differentially expressed, the calibration is done by labeling samples of the calibrating cDNA with the two fluorophores and adding identical amounts of each to the hybridization mixture.

The output of the photomultiplier tube is digitized using a 12-bit RTI-835H analog-to-digital (A/D) conversion board (Analog Devices, Inc., Norwood MA) installed in an IBM-compatible PC computer. The digitized data are displayed as an image where the signal intensity is mapped using a linear 20-color transformation to a pseudocolor scale ranging from blue (low signal) to red (high signal). The data is also analyzed quantitatively. Where two different fluorophores are excited and measured simultaneously, the data are first corrected for optical crosstalk (due to overlapping emission spectra) between the fluorophores using each fluorophore's emission spectrum.

A grid is superimposed over the fluorescence signal image such that the signal from each spot is centered in each element of the grid. The fluorescence signal within each element is then integrated

to obtain a numerical value corresponding to the average intensity of the signal. The software used for signal analysis is the GEMTOOLS gene expression analysis program (Incyte). Array elements that exhibit at least about a two-fold change in expression, a signal-to-background ratio of at least about 2.5, and an element spot size of at least about 40%, are considered to be differentially expressed.

Expression

SEQ ID NO:43 showed differential expression, as determined by microarray analysis. For example, SEQ ID NO:43 showed differential expression in treated versus untreated Jurkat cells, as determined by microarray analysis. Array elements that exhibited about at least a two-fold change in expression, a signal-to-background ratio of a least 2.5, and an element spot size of at least 40% were identified as differentially expressed using the GEMTOOLS program (Incyte Genomics).

In an alternative example, expression of SEQ ID NO:43 was down regulated in PMA plus ionomycin-treated Jurkat cells versus untreated Jurkat cells as determined by microarray analysis. Jurkat cells were treated with combinations of graded doses of PMA and ionomycin and collected at a 1 hour time point. The treated cells were compared to untreated Jurkat cells kept in culture in the absence of stimuli.

In similar experiments, expression of SEQ ID NO:43 was down regulated in Jurkat cells stimulated *in vitro* with 1μg soluble mouse anti-human CD3 and compared to untreated Jurkat cells kept in culture in the absence of stimuli. Differential expression was significant in the cells treated for 1, 2, and 4 hours; the results at 8 hours were not statistically significant.

In an alternative example, PHA blasts were derived from the PBMCs of 5 healthy volunteer donors. The PBMCs were stimulated for 12 days in presence of PHA and IL-2. These T cell blasts were washed and stimulated for 2 hours in the presence of anti-CD3 monoclonal antibody, anti-CD28 antibody, a combination of both antibodies, PMA, ionomycin, and a combination of PMA and ionomycin. These reactivated T cells were compared to matching untreated PHA blasts. SEQ ID NO:78 was found to be downregulated by at least two-fold in cells stimulated in the presence of anti-CD3 + PMA, anti-CD3 + anti-CD28, PMA + ionomycin, and PMA alone in the one donor tested. Therefore, in various embodiments, SEQ ID NO:43 and SEQ ID NO:78 can be used for one or more of the following: i) monitoring treatment of immune disorders and related diseases and conditions, ii) diagnostic assays for immune disorders and related diseases and conditions, and iii) developing therapeutics and/or other treatments for immune disorders and related diseases and conditions.

Expression of SEQ ID NO:44, SEQ ID NO:46, and SEQ ID NO:78 showed differential expression in tumorous or diseased colon tissue versus non-tumorous or healthy colon tissues, as

determined by microarray analysis. Array elements that exhibited about at least a two-fold change in expression, a signal-to-background ratio of a least 2.5, and an element spot size of at least 40% were identified as differentially expressed using the GEMTOOLS program (Incyte Genomics). SEQ ID NO:44 exhibited at least a two-fold decrease in colon polyps, and at least a two-fold increase in sigmoidal colon sarcoma tissue. SEQ ID NO:46 exhibited upregulation in colon adenocarcinoma tissue, and in sigmoidal colon sarcoma tissue. SEQ ID NO:78 was downregulated by at least two-fold in matched normal versus tumorous colon tissues in one out of thirteen donors tested. Therefore, in various embodiments, SEQ ID NO:44, SEQ ID NO:46, and SEQ ID NO:78 can be used for one or more of the following: i) monitoring treatment of colon cancer, ii) diagnostic assays for colon cancer, and iii) developing therapeutics and/or other treatments for colon cancer.

SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:52, SEQ ID NO:53, and SEQ ID NO:83 showed differential expression in breast cancer cell lines, as determined by microarray analysis. The gene expression profile of a nonmalignant mammary epithelial cell line (HMEC) was compared to the gene expression profiles of breast carcinoma lines at different stages of tumor progression. Cell lines compared included: a) BT-20, a breast carcinoma cell line derived *in vitro* from the cells emigrating out of thin slices of tumor mass isolated from a 74-year-old female, b) BT-474, a breast ductal carcinoma cell line that was isolated from a solid, invasive ductal carcinoma of the breast obtained from a 60-year-old woman, c) BT-483, a breast ductal carcinoma cell line that was isolated from a papillary invasive ductal tumor obtained from a 23-year-old normal, menstruating, parous female with a family history of breast cancer, d) Hs 578T, a breast ductal carcinoma cell line isolated from a 74-year-old female with breast carcinoma, e) MCF7, a nonmalignant breast adenocarcinoma cell line isolated from the pleural effusion of a 69-year-old female, f) MCF-10A, a breast mammary gland (luminal ductal characteristics) cell line isolated from a 36-year-old woman with fibrocystic breast disease, g) MDA-MB-468, a breast adenocarcinoma cell line isolated from the pleural effusion of a 51-year-old female with metastatic adenocarcinoma of the breast, and h) HMEC, primary breast epithelial cells isolated from a normal donor. Expression of SEQ ID NO:48 was increased at least 2-fold in the Hs 578T cell line when cultured under optimal growth conditions or starved, when compared to expression levels detected in starved HMECs. Expression of SEQ ID NO:49 was decreased at least 2-fold in BT-474 cells grown under optimal conditions, at least 2-fold in starved Hs 578T cells, at least 2.5-fold in BT-483 cells grown under optimal conditions, and at least 3.4-fold in starved MCF7 cells, when compared to expression levels in starved HMECs. Expression of SEQ ID NO:53 was increased at least two-fold in a breast carcinoma cell line (Hs 578T) grown in mammary epithelium growth medium (MEGM) or under starvation conditions versus HMECs grown under

starvation conditions.

Further, SEQ ID NO:52 was down-regulated in several breast cancer cell lines versus a primary cell culture of normal mammary epithelial cells. The gene expression profile of a nonmalignant mammary epithelial cell line was compared to the gene expression profiles of breast carcinoma lines at different stages of tumor progression. Expression of SEQ ID NO:52 was decreased at least 2.5-fold in four breast carcinoma cell lines (BT-20, BT-474, BT-483, and MCF7) grown in mammary epithelium growth medium (MEGM) or under starvation conditions versus HMECs grown under starvation conditions. Expression of SEQ ID NO:52 was increased at least 3-fold in breast cancer cell line Hs 578T grown in MEGM versus HMECs grown under both starvation conditions and MEGM. Although expression of SEQ ID NO:52 was not affected in the same manner among all breast cancer cell lines, the data suggest that in some populations or stages of breast cancer this protein is differentially expressed and thus might provide a useful screening or monitoring tool for breast cancer. Further, SEQ ID NO:52 was up-regulated in several breast carcinoma cell lines when compared with non-tumorigenic mammary cells (MCF10A) from a donor with fibrocystic disease. The gene expression profile of a nonmalignant mammary epithelial cell line was compared to the gene expression profiles of breast carcinoma lines at different stages of tumor progression. Cell lines compared included: a) MCF-10A (see above); b)MCF7 (see above); c)T-47D, a breast carcinoma cell line isolated from a pleural effusion obtained from a 54-year-old female with an infiltrating ductal carcinoma of the breast; d)Sk-BR-3, a breast adenocarcinoma cell line isolated from a malignant pleural effusion of a 43-year-old female; e)BT-20 (see above); f)MDA-mb-231, a breast tumor cell line isolated from the pleural effusion of a 51-year old female; and g) MDA-mb-435S, a spindle shaped strain that evolved from the parent line (435) isolated from the pleural effusion of a 31-year-old female with metastatic, ductal adenocarcinoma of the breast. Expression of SEQ ID NO:52 was increased from 2- to 8-fold in untreated breast cancer cell lines BT-20 and MDAM231 versus untreated non-tumorigenic mammary cells (MCF10A).

Further, SEQ ID NO:83 showed differential expression, as determined by microarray analysis. For example, expression of SEQ ID NO:83 was down-regulated in human breast cancer cell lines (ductal carcinoma and adenocarcinoma) versus normal human mamary epithelial cells (HMEC). Expression of SEQ ID NO:83 was decreased at least two-fold in all breast cancer cell line s evaluated, with the exception of one cell line originally isolated from a patient with nonmalignant, nontumorigenic fibrocystic disease. Therefore, in various embodiments, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:52, SEQ ID NO:53, and SEQ ID NO:83 can be used for one or more of the following: i) monitoring treatment of breast cancer, ii) diagnostic assays for breast cancer, and iii)

developing therapeutics and/or other treatments for breast cancer.

In another example, SEQ ID NO:48 and SEQ ID NO:52 showed differential expression in prostate cancer cell lines, as determined by microarray expression analysis. Primary prostate epithelial cells were compared with prostate carcinomas representative of the different stages of tumor progression. Cell lines compared included: a) PrEC, a primary prostate epithelial cell line isolated from a normal donor, b) DU 145, a prostate carcinoma cell line isolated from a metastatic site in the brain of 69-year old male with widespread metastatic prostate carcinoma, c) LNCaP, a prostate carcinoma cell line isolated from a lymph node biopsy of a 50-year-old male with metastatic prostate carcinoma, and d) PC-3, a prostate adenocarcinoma cell line isolated from a metastatic site in the bone of a 62-year-old male with grade IV prostate adenocarcinoma. In one example, SEQ ID NO:48 expression was decreased at least 2.5-fold in PC-3 cells grown in basal media in the absence of growth factors and hormones, when compared to normal PrECs grown under the same conditions. In another example, SEQ ID NO:48 expression was decreased at least 2-fold in LNCaP and DU 145 cells grown under optimal growth conditions, in the presence of growth factors and nutrients, when compared to normal PrECs grown under the same conditions. Expression of SEQ ID NO:52 was decreased from 2.5- to 7-fold in two prostate cancer cell lines (PC-3 and LNCaP) when grown under restrictive (basal media in the absence of growth factors and hormones) or optimal (presence of growth factors and nutrients) versus PrECs grown under restrictive conditions. Therefore, in various embodiments, SEQ ID NO:48 and SEQ ID NO:52 can be used for one or more of the following: i) monitoring treatment of prostate cancer, ii) diagnostic assays for prostate cancer, and iii) developing therapeutics and/or other treatments for prostate cancer.

In another example, SEQ ID NO:50 was differentially expressed in adipocytes isolated from an obese donor, as determined by microarray expression analysis. Primary subcutaneous preadipocytes were isolated from adipose tissue of a 28-year-old healthy female with body mass index (BMI) of 23.59 (normal donor), and from adipose tissue of a 40-year-old healthy female with a body mass index (BMI) of 32.47 (obese donor). The preadipocytes were cultured and induced to differentiate into adipocytes by culturing them in differentiation medium containing active components PPAR-γ agonist and human insulin (Zen-Bio). Thiazolidinediones or PPAR-γ agonists can bind and activate an orphan nuclear receptor, PPAR-γ, and some of them have been proven to be able to induce human adipocyte differentiation. The preadipocytes were treated with human insulin and PPAR-γ agonist for 3 days and subsequently were switched to medium containing insulin for a variety of time periods ranging from one to 20 days before the cells were collected for analysis. Differentiated adipocytes were compared to untreated preadipocytes maintained in culture in the

absence of inducing agents. Between 80% and 90% of the preadipocytes finally differentiated to adipocytes as observed under phase contrast microscope. Expression levels of SEQ ID NO:50 decreased at least 2-fold after 48 hours of treatment with differentiation media in the preadipocytes from the obese donor, when compared to untreated cells from the same donor. The decrease in expression of SEQ ID NO:50 peaked at approximately 3.4-fold after 1.1 week, and continued to be at least 2-fold through 2.1 weeks of culture in the differentiation media. This decrease in SEQ ID NO:50 expression was not seen in the preadipocytes isolated from the normal donor upon culture in the differentiation media. Therefore, in various embodiments, SEQ ID NO:50 can be used for one or more of the following: i) monitoring treatment of diabetes mellitus and other, obesity-related disorders, ii) diagnostic assays for diabetes mellitus and other, obesity-related disorders, and iii) developing therapeutics and/or other treatments for diabetes mellitus and other, obesity-related disorders.

In another example, SEQ ID NO:52 was up-regulated in ovarian adenocarcinoma versus normal ovarian tissue from the same donor as determined by microarray analysis. A normal ovary from a 79 year-old female donor was compared to an ovarian adenocarcinoma from the same donor (Huntsman Cancer Institute, Salt Lake City, UT). Expression of SEQ ID NO:52 was increased at least two-fold in the ovarian adenocarcinoma tissue as compared to normal ovarian tissue from the same donor. Therefore, in various embodiments, SEQ ID NO:52 can be used for one or more of the following: i) monitoring treatment of ovarian cancer, ii) diagnostic assays for ovarian cancer, and iii) developing therapeutics and/or other treatments for ovarian cancer.

In another example, SEQ ID NO:52 was up-regulated in fibroblasts from a patient with Tangier disease versus fibroblasts from a normal subject as determined by microarray analysis. Normal and Tangier disease derived fibroblasts were compared. Human fibroblasts were obtained from skin explants from both normal subjects and two patients homozygous for Tangier disease. Cell lines were immortalized by transfection with human papillomavirus 16 genes E6 and E7 and a neomycin resistance selectable marker. In addition, both types of cells were cultured in the presence of cholesterol and compared with the same cell type cultured in the absence of cholesterol. TD derived cells are shown to be deficient in an assay of apoA-I mediated tritiated cholesterol efflux. Expression of SEQ ID NO:52 was increased at least five-fold in fibroblasts from a patient with Tangier disease versus fibroblasts from a normal subject. Further, SEQ ID NO:78 was downregulated by at least two-fold in both types of comparisons and SEQ ID NO:80 was downregulated by at least two-fold in Tangier disease derived fibroblasts cultured in the presence of cholesterol when compared with normal fibroblasts cultured in the presence of cholesterol. Therefore, in various embodiments, SEQ ID NO:52, SEQ ID NO:78, and/or SEQ ID NO:80 can be used for one

or more of the following: i) monitoring treatment of Tangier disease, ii) diagnostic assays for Tangier disease, and iii) developing therapeutics and/or other treatments for Tangier disease.

In another example, SEQ ID NO:52 was up-regulated in a spontaneously transformed endothelial cell line (ECV304) treated with TNF-α versus untreated ECV304 cells as determined by microarray analysis. ECV304 cells were grown to 85% confluency and then treated with a titration of concentations of TNF-α for 0, 1, 2, 8, and 24 hours. TNF-α is produced by activated lymphocytes, macrophages, and other white blood cells and can activate endothelial cells. Monitoring the endothelial cells' response to TNF-α at the level of mRNA expression can provide information necessary for better understanding of both TNF-α signaling pathways and endothelial cell biology. Expression of SEQ ID NO:52 was increased at least two-fold in ECV304 cells treated for two hours with varying concentrations of TNF-α as compared to untreated ECV304 cells. Therefore, in various embodiments, SEQ ID NO:52 can be used for one or more of the following: i) monitoring treatment of inflammation, vascular disease, and related diseases and conditions, ii) diagnostic assays for inflammation, vascular disease, and related diseases and conditions, and iii) developing therapeutics and/or other treatments for inflammation, vascular disease, related diseases and conditions.

SEQ ID NO:69 and SEQ ID NO:73 showed at least a two-fold decrease in expression in C3A cells treated with gemfibrozil compared to untreated cells as determined by microarray analysis. C3A cells were treated with 120, 600, 800 or 1200 μM gemfibrozil for 1, 3 or 6 hours. Therefore, in various embodiments, SEQ ID NO:69 and SEQ ID NO:73 can each be used for one or more of the following: i) monitoring treatment of coronary heart disease, hyperlipoproteinemia, obesity, gall bladder disease, stroke, and hyperlipidemia, ii) diagnostic assays for coronary heart disease, hyperlipoproteinemia, obesity, gall bladder disease, stroke, and hyperlipidemia, and iii) developing therapeutics and/or other treatments for coronary heart disease, hyperlipoproteinemia, obesity, gall bladder disease, stroke, and hyperlipidemia.

In an alternative example, SEQ ID NO:70 and SEQ ID NO:78 showed differential expression associated with lung cancer. Expression in tumorous tissue from ten patients with lung cancer was compared to grossly uninvolved lung tissue from the same donors. SEQ ID NO:70 showed at least a two-fold decrease in expression in lung tissue from three out of five patients with squamous cell cancer compared to matched microscopically normal tissue from the same donors as determined by microarray analysis. Further, SEQ ID NO:78 was downregulated by at least two-fold in matched normal versus tumorous lung tissues in three out of seven donors tested. Therefore, in various embodiments, SEQ ID NO:70 and SEQ ID NO:78 can be used for one or more of the following: i) monitoring treatment of lung cancer, ii) diagnostic assays for lung cancer, and iii) developing

therapeutics and/or other treatments for lung cancer.

In another example, SEQ ID NO:57 showed tissue-specific expression as determined by microarray analysis. RNA samples isolated from a variety of normal human tissues were compared to a common reference sample. Tissues contributing to the reference sample were selected for their ability to provide a complete distribution of RNA in the human body and include brain (4%), heart (7%), kidney (3%), lung (8%), placenta (46%), small intestine (9%), spleen (3%), stomach (6%), testis (9%), and uterus (5%). The normal tissues assayed were obtained from at least three different donors. RNA from each donor was separately isolated and individually hybridized to the microarray. Since these hybridization experiments were conducted using a common reference sample, differential expression values are directly comparable from one tissue to another. The expression of SEQ ID NO:57 was increased by at least two-fold in pancreatic tissue as compared to the reference sample. Therefore, SEQ ID NO:57 can be used as a marker for pancreatic tissue.

In an alternative example, SEQ ID NO:66 showed tissue-specific expression as determined by microarray analysis. RNA samples isolated from a variety of normal human tissues were compared to a common reference sample. Tissues contributing to the reference sample were selected for their ability to provide a complete distribution of RNA in the human body and include brain (4%), heart (7%), kidney (3%), lung (8%), placenta (46%), small intestine (9%), spleen (3%), stomach (6%), testis (9%), and uterus (5%). The normal tissues assayed were obtained from at least three different donors. RNA from each donor was separately isolated and individually hybridized to the microarray. Since these hybridization experiments were conducted using a common reference sample, differential expression values are directly comparable from one tissue to another. The expression of SEQ ID NO:66 was increased by at least two-fold in kidney as compared to the reference sample. Therefore, SEQ ID NO:66 can be used as a tissue marker for kidney.

In a further example, SEQ ID NO:70 showed tissue-specific expression as determined by microarray analysis. RNA samples isolated from a variety of normal human tissues were compared to a common reference sample. Tissues contributing to the reference sample were selected for their ability to provide a complete distribution of RNA in the human body and include brain (4%), heart (7%), kidney (3%), lung (8%), placenta (46%), small intestine (9%), spleen (3%), stomach (6%), testis (9%), and uterus (5%). The normal tissues assayed were obtained from at least three different donors. RNA from each donor was separately isolated and individually hybridized to the microarray. Since these hybridization experiments were conducted using a common reference sample, differential expression values are directly comparable from one tissue to another. The expression of SEQ ID NO:70 was increased by at least two-fold in omentum as compared to the reference sample.

Therefore, SEQ ID NO:70 can be used as a tissue marker for omentum.

XII. Complementary Polynucleotides

Sequences complementary to the CADECM-encoding sequences, or any parts thereof, are used to detect, decrease, or inhibit expression of naturally occurring CADECM. Although use of oligonucleotides comprising from about 15 to 30 base pairs is described, essentially the same procedure is used with smaller or with larger sequence fragments. Appropriate oligonucleotides are designed using OLIGO 4.06 software (National Biosciences) and the coding sequence of CADECM. To inhibit transcription, a complementary oligonucleotide is designed from the most unique 5' sequence and used to prevent promoter binding to the coding sequence. To inhibit translation, a complementary oligonucleotide is designed to prevent ribosomal binding to the CADECM-encoding transcript.

XIII. Expression of CADECM

Expression and purification of CADECM is achieved using bacterial or virus-based expression systems. For expression of CADECM in bacteria, cDNA is subcloned into an appropriate vector containing an antibiotic resistance gene and an inducible promoter that directs high levels of cDNA transcription. Examples of such promoters include, but are not limited to, the *trp-lac* (*tac*) hybrid promoter and the T5 or T7 bacteriophage promoter in conjunction with the *lac* operator regulatory element. Recombinant vectors are transformed into suitable bacterial hosts, e.g., BL21(DE3). Antibiotic resistant bacteria express CADECM upon induction with isopropyl beta-D-thiogalactopyranoside (IPTG). Expression of CADECM in eukaryotic cells is achieved by infecting insect or mammalian cell lines with recombinant *Autographica californica* nuclear polyhedrosis virus (AcMNPV), commonly known as baculovirus. The nonessential polyhedrin gene of baculovirus is replaced with cDNA encoding CADECM by either homologous recombination or bacterial-mediated transposition involving transfer plasmid intermediates. Viral infectivity is maintained and the strong polyhedrin promoter drives high levels of cDNA transcription. Recombinant baculovirus is used to infect *Spodoptera frugiperda* (Sf9) insect cells in most cases, or human hepatocytes, in some cases. Infection of the latter requires additional genetic modifications to baculovirus (Engelhard, E.K. et al. (1994) Proc. Natl. Acad. Sci. USA 91:3224-3227; Sandig, V. et al. (1996) Hum. Gene Ther. 7:1937-1945).

In most expression systems, CADECM is synthesized as a fusion protein with, e.g., glutathione S-transferase (GST) or a peptide epitope tag, such as FLAG or 6-His, permitting rapid, single-step, affinity-based purification of recombinant fusion protein from crude cell lysates. GST, a 26-kilodalton enzyme from *Schistosoma japonicum*, enables the purification of fusion proteins on immobilized glutathione under conditions that maintain protein activity and antigenicity (Amersham

Biosciences). Following purification, the GST moiety can be proteolytically cleaved from CADECM at specifically engineered sites. FLAG, an 8-amino acid peptide, enables immunoaffinity purification using commercially available monoclonal and polyclonal anti-FLAG antibodies (Eastman Kodak). 6-His, a stretch of six consecutive histidine residues, enables purification on metal-chelate resins (QIAGEN). Methods for protein expression and purification are discussed in Ausubel et al. (*supra*, ch. 10 and 16). Purified CADECM obtained by these methods can be used directly in the assays shown in Examples XVII and XVIII, where applicable.

XIV. Functional Assays

CADECM function is assessed by expressing the sequences encoding CADECM at physiologically elevated levels in mammalian cell culture systems. cDNA is subcloned into a mammalian expression vector containing a strong promoter that drives high levels of cDNA expression. Vectors of choice include PCMV SPORT plasmid (Invitrogen, Carlsbad CA) and PCR3.1 plasmid (Invitrogen), both of which contain the cytomegalovirus promoter. 5-10 μg of recombinant vector are transiently transfected into a human cell line, for example, an endothelial or hematopoietic cell line, using either liposome formulations or electroporation. 1-2 μg of an additional plasmid containing sequences encoding a marker protein are co-transfected. Expression of a marker protein provides a means to distinguish transfected cells from nontransfected cells and is a reliable predictor of cDNA expression from the recombinant vector. Marker proteins of choice include, e.g., Green Fluorescent Protein (GFP; BD Clontech), CD64, or a CD64-GFP fusion protein. Flow cytometry (FCM), an automated, laser optics-based technique, is used to identify transfected cells expressing GFP or CD64-GFP and to evaluate the apoptotic state of the cells and other cellular properties. FCM detects and quantifies the uptake of fluorescent molecules that diagnose events preceding or coincident with cell death. These events include changes in nuclear DNA content as measured by staining of DNA with propidium iodide; changes in cell size and granularity as measured by forward light scatter and 90 degree side light scatter; down-regulation of DNA synthesis as measured by decrease in bromodeoxyuridine uptake; alterations in expression of cell surface and intracellular proteins as measured by reactivity with specific antibodies; and alterations in plasma membrane composition as measured by the binding of fluorescein-conjugated Annexin V protein to the cell surface. Methods in flow cytometry are discussed in Ormerod, M.G. (1994; Flow Cytometry, Oxford, New York NY).

The influence of CADECM on gene expression can be assessed using highly purified populations of cells transfected with sequences encoding CADECM and either CD64 or CD64-GFP. CD64 and CD64-GFP are expressed on the surface of transfected cells and bind to conserved regions

of human immunoglobulin G (IgG). Transfected cells are efficiently separated from nontransfected cells using magnetic beads coated with either human IgG or antibody against CD64 (DYNAL, Lake Success NY). mRNA can be purified from the cells using methods well known by those of skill in the art. Expression of mRNA encoding CADECM and other genes of interest can be analyzed by northern analysis or microarray techniques.

XV. Production of CADECM Specific Antibodies

CADECM substantially purified using polyacrylamide gel electrophoresis (PAGE; see, e.g., Harrington, M.G. (1990) Methods Enzymol. 182:488-495), or other purification techniques, is used to immunize animals (e.g., rabbits, mice, etc.) and to produce antibodies using standard protocols.

Alternatively, the CADECM amino acid sequence is analyzed using LASERGENE software (DNASTAR) to determine regions of high immunogenicity, and a corresponding oligopeptide is synthesized and used to raise antibodies by means known to those of skill in the art. Methods for selection of appropriate epitopes, such as those near the C-terminus or in hydrophilic regions are well described in the art (Ausubel et al., *supra*, ch. 11).

Typically, oligopeptides of about 15 residues in length are synthesized using an ABI 431A peptide synthesizer (Applied Biosystems) using FMOC chemistry and coupled to KLH (Sigma-Aldrich, St. Louis MO) by reaction with N-maleimidobenzoyl-N-hydroxysuccinimide ester (MBS) to increase immunogenicity (Ausubel et al., *supra*). Rabbits are immunized with the oligopeptide-KLH complex in complete Freund's adjuvant. Resulting antisera are tested for antipeptide and anti-CADECM activity by, for example, binding the peptide or CADECM to a substrate, blocking with 1% BSA, reacting with rabbit antisera, washing, and reacting with radio-iodinated goat anti-rabbit IgG.

XVI. Purification of Naturally Occurring CADECM Using Specific Antibodies

Naturally occurring or recombinant CADECM is substantially purified by immunoaffinity chromatography using antibodies specific for CADECM. An immunoaffinity column is constructed by covalently coupling anti-CADECM antibody to an activated chromatographic resin, such as CNBr-activated SEPHAROSE (Amersham Biosciences). After the coupling, the resin is blocked and washed according to the manufacturer's instructions.

Media containing CADECM are passed over the immunoaffinity column, and the column is washed under conditions that allow the preferential absorbance of CADECM (e.g., high ionic strength buffers in the presence of detergent). The column is eluted under conditions that disrupt antibody/CADECM binding (e.g., a buffer of pH 2 to pH 3, or a high concentration of a chaotrope, such as urea or thiocyanate ion), and CADECM is collected.

XVII. Identification of Molecules Which Interact with CADECM

CADECM, or biologically active fragments thereof, are labeled with ^{125}I Bolton-Hunter reagent (Bolton, A.E. and W.M. Hunter (1973) Biochem. J. 133:529-539). Candidate molecules previously arrayed in the wells of a multi-well plate are incubated with the labeled CADECM, washed, and any wells with labeled CADECM complex are assayed. Data obtained using different concentrations of CADECM are used to calculate values for the number, affinity, and association of CADECM with the candidate molecules.

Alternatively, molecules interacting with CADECM are analyzed using the yeast two-hybrid system as described in Fields, S. and O. Song (1989; Nature 340:245-246), or using commercially available kits based on the two-hybrid system, such as the MATCHMAKER system (BD Clontech).

CADECM may also be used in the PATHCALLING process (CuraGen Corp., New Haven CT) which employs the yeast two-hybrid system in a high-throughput manner to determine all interactions between the proteins encoded by two large libraries of genes (Nandabalan, K. et al. (2000) U.S. Patent No. 6,057,101).

XVIII. Demonstration of CADECM Activity

An assay for CADECM activity measures the expression of CADECM on the cell surface. cDNA encoding CADECM is transfected into a non-leukocytic cell line. Cell surface proteins are labeled with biotin (de la Fuente, M.A. et al. (1997) Blood 90:2398-2405). Immunoprecipitations are performed using CADECM-specific antibodies, and immunoprecipitated samples are analyzed using SDS-PAGE and immunoblotting techniques. The ratio of labeled immunoprecipitant to unlabeled immunoprecipitant is proportional to the amount of CADECM expressed on the cell surface.

Alternatively, an assay for CADECM activity measures the amount of cell aggregation induced by overexpression of CADECM. In this assay, cultured cells such as NIH3T3 are transfected with cDNA encoding CADECM contained within a suitable mammalian expression vector under control of a strong promoter. Cotransfection with cDNA encoding a fluorescent marker protein, such as Green Fluorescent Protein (CLONTECH), is useful for identifying stable transfectants. The amount of cell agglutination, or clumping, associated with transfected cells is compared with that associated with untransfected cells. The amount of cell agglutination is a direct measure of CADECM activity.

Alternatively, an assay for CADECM activity measures the disruption of cytoskeletal filament networks upon overexpression of CADECM in cultured cell lines (Rezniczek, G. A. et al. (1998) J. Cell Biol. 141:209-225). cDNA encoding CADECM is subcloned into a mammalian expression vector that drives high levels of cDNA expression. This construct is transfected into cultured cells, such as rat kangaroo PtK2 or rat bladder carcinoma 804G cells. Actin filaments and intermediate filaments

such as keratin and vimentin are visualized by immunofluorescence microscopy using antibodies and techniques well known in the art. The configuration and abundance of cyoskeletal filaments can be assessed and quantified using confocal imaging techniques. In particular, the bundling and collapse of cytoskeletal filament networks is indicative of CADECM activity.

Alternatively, cell adhesion activity in CADECM is measured in a 96-well plate in which wells are first coated with CADECM by adding solutions of CADECM of varying concentrations to the wells. Excess CADECM is washed off with saline, and the wells incubated with a solution of 1% bovine serum albumin to block non-specific cell binding. Aliquots of a cell suspension of a suitable cell type are then added to the wells and incubated for a period of time at 37 °C. Non-adherent cells are washed off with saline and the cells stained with a suitable cell stain such as Coomassie blue. The intensity of staining is measured using a variable wavelength multi-well plate reader and compared to a standard curve to determine the number of cells adhering to the CADECM coated plates. The degree of cell staining is proportional to the cell adhesion activity of CADECM in the sample.

Various modifications and variations of the described compositions, methods, and systems of the invention will be apparent to those skilled in the art without departing from the scope and spirit of the invention. It will be appreciated that the invention provides novel and useful proteins, and their encoding polynucleotides, which can be used in the drug discovery process, as well as methods for using these compositions for the detection, diagnosis, and treatment of diseases and conditions. Although the invention has been described in connection with certain embodiments, it should be understood that the invention as claimed should not be unduly limited to such specific embodiments. Nor should the description of such embodiments be considered exhaustive or limit the invention to the precise forms disclosed. Furthermore, elements from one embodiment can be readily recombined with elements from one or more other embodiments. Such combinations can form a number of embodiments within the scope of the invention. It is intended that the scope of the invention be defined by the following claims and their equivalents.

Table 1

Incyte Project ID	Polypeptide SEQ ID NO:	Incyte Polypeptide ID	Polynucleotide SEQ ID NO:	Incyte Polynucleotide ID	Incyte Full Length Clones
7513225	1	7513225CD1	43	7513225CB1	
7513288	2	7513288CD1	44	7513288CB1	
7513607	3	7513607CD1	45	7513607CB1	90035803CA2
7513991	4	7513991CD1	46	7513991CB1	95013812CA2
7513298	5	7513298CD1	47	7513298CB1	
7517764	6	7517764CD1	48	7517764CB1	
7517774	7	7517774CD1	49	7517774CB1	90127706CA2
7518133	8	7518133CD1	50	7518133CB1	
7520147	9	7520147CD1	51	7520147CB1	95114711CA2
7520276	10	7520276CD1	52	7520276CB1	95090890CA2
7520808	11	7520808CD1	53	7520808CB1	95119080CA2
7520821	12	7520821CD1	54	7520821CB1	3808913CA2, 3982690CA2, 4804040CA2, 5719874CA2, 6598743CA2
7520839	13	7520839CD1	55	7520839CB1	95114992CA2, 95115004CA2, 95115128CA2
7520891	14	7520891CD1	56	7520891CB1	95114929CA2, 95115002CA2, 95115010CA2, 95115052CA2, 95115092CA2, 95115104CA2, 95115168CA2, 95115192CA2
7514645	15	7514645CD1	57	7514645CB1	90205460CA2, 90205484CA2, 90208373CA2, 90208389CA2, 90208433CA2, 90208441CA2, 90208525CA2, 90208633CA2
7517776	16	7517776CD1	58	7517776CB1	
7517783	17	7517783CD1	59	7517783CB1	
7522607	18	7522607CD1	60	7522607CB1	90125404CA2
7521142	19	7521142CD1	61	7521142CB1	95119390CA2
7521689	20	7521689CD1	62	7521689CB1	
2878775	21	2878775CD1	63	2878775CB1	
7521207	22	7521207CD1	64	7521207CB1	95125625CA2, 95125641CA2, 95125717CA2, 95125733CA2, 95125741CA2
7521283	23	7521283CD1	65	7521283CB1	95115020CA2, 95115028CA2, 95115029CA2, 95115036CA2, 95125609CA2
7522210	24	7522210CD1	66	7522210CB1	90072007CA2

Table 1

Incyte Project ID	Polypeptide SEQ ID NO:	Incyte Polypeptide ID	Polynucleotide SEQ ID NO:	Incyte Polynucleotide ID	Incyte Full Length Clones
7519488	25	7519488CD1	67	7519488CB1	95081815CA2, 95082070CA2, 95103323CA2, 95103486CA2
7519965	26	7519965CD1	68	7519965CB1	95103287CA2
7519985	27	7519985CD1	69	7519985CB1	95103283CA2, 95103559CA2
7520002	28	7520002CD1	70	7520002CB1	95114424CA2, 95114532CA2
7520014	29	7520014CD1	71	7520014CB1	95114435CA2, 95114491CA2
7520039	30	7520039CD1	72	7520039CB1	95126975CA2
7520053	31	7520053CD1	73	7520053CB1	95103519CA2, 95103527CA2
7523262	32	7523262CD1	74	7523262CB1	95169993CA2, 95170053CA2, 95182601CA2
7523270	33	7523270CD1	75	7523270CB1	95167922CA2
7523287	34	7523287CD1	76	7523287CB1	95170218CA2
7521825	35	7521825CD1	77	7521825CB1	95147859CA2
7521844	36	7521844CD1	78	7521844CB1	95139693CA2
7521864	37	7521864CD1	79	7521864CB1	95137780CA2, 95137824CA2, 95137888CA2
7522020	38	7522020CD1	80	7522020CB1	95145645CA2
758410	39	758410CD1	81	758410CB1	
7520759	40	7520759CD1	82	7520759CB1	95121966CA2
7522915	41	7522915CD1	83	7522915CB1	95064105CA2
7522936	42	7522936CD1	84	7522936CB1	95082130CA2

Table 2

Polypeptide SEQ ID NO:	Incyte Polypeptide ID	GenBank ID NO: or PROTEOME ID NO:	Probability Score	Annotation
1	7513225CD1	g2791962	0.0	[Homo sapiens] nidogen-2
				Kohfeldt, E. et al. Nidogen-2: a new basement membrane protein with diverse binding properties. J. Mol. Biol. 282, 99-109 (1998).
		429052|NID2	0.0	[Homo sapiens][Extracellular matrix (cuticle and basement membrane); Basement membrane (extracellular matrix); Extracellular (excluding cell wall)] Nidogen 2, a basement membrane protein that binds perlecan (HSPG2), laminin-1, and collagens; supports cell attachment in some cell lines
				Hopf, M. et al. Recombinant domain IV of perlecan binds to nidogens, laminin-nidogen complex, fibronectin, fibulin-2 and heparin. Eur. J. Biochem. 259, 917-925 (1999).
		582237|Nid2	0.0	[Mus musculus][Extracellular matrix (cuticle and basement membrane); Basement membrane (extracellular matrix); Extracellular (excluding cell wall)] Nidogen 2, a basement membrane protein that localizes to elastic fibers of blood vessel walls; has integrin-binding RGD sequence; may facilitate stabilization of basement membrane networks
				Kimura, N. et al. Entactin-2: a new member of basement membrane protein with high homology to entactin/nidogen. Exp. Cell Res. 241, 36-45 (1998).
2	7513288CD1	g556845	0.0	[Homo sapiens] human tenascin-C
				Gherzi, R. et al. Human tenascin gene. Structure of the 5'-region, identification, and characterization of the transcription regulatory sequences, J. Biol. Chem. 270, 3429-3434 (1995).
		335910|HXB	0.0	[Homo sapiens] [Adhesin/agglutinin; Receptor (signalling)] [Extracellular matrix (cuticle and basement membrane); Extracellular (excluding cell wall)] Tenascin C (hexabrachion), a multidomain extracellular matrix glycoprotein, may be involved in cell adhesion and neurite outgrowth, induced by hypoxia and has increased expression in colorectal carcinoma, osteosarcoma metastases and pulmonary hypertension
				Talts, J. F. et al. Regulation of mesenchymal extracellular matrix protein synthesis by transforming growth factor-beta and glucocorticoids in tumor stroma. J. Cell Sci. 2153-2162 (1995).

Table 2

Polypeptide SEQ ID NO:	Incyte Polypeptide ID	GenBank ID NO: or PROTEOME ID NO:	Probability Score	Annotation
		586659|Tnc	0.0	[Mus musculus] [Structural protein] [Extracellular matrix (cuticle and basement membrane); Extracellular (excluding cell wall)] Tenascin C (hexabrachion), a multidomain extracellular matrix glycoprotein with epidermal growth factor-like and fibronectin type III-like repeats, may be involved in cell adhesion, has effects on cell proliferation and
				Wenk, M. B. et al. Tenascin-C suppresses Rho activation. J. Cell Biol. 150, 913-920
3	7513607CD1	g31419	1.3E-29	[Homo sapiens] fibulin-1 C
				Argraves, W. S. et al. Fibulin is an extracellular matrix and plasma glycoprotein with repeated domain structure J. Cell Biol. 111 (6 Pt 2), 3155-3164 (1990).
		582841|Fbln2	1.9E-24	[Mus musculus] [Structural protein] [Extracellular matrix (cuticle and basement membrane); Basement membrane (extracellular matrix); Extracellular (excluding cell wall)] Fibulin 2, a putative extracellular matrix protein that plays a role in cell adhesion and may play a role in hemostatic control, angiogenesis, cardiogenesis, and maintenance of the heart septa and valve structure
				Zhang, R. Z. et al. Fibulin-2 (FBLN2): human cDNA sequence, mRNA expression, and mapping of the gene on human and mouse chromosomes. Genomics 22, 425-430 (1994).
		339796|FBLN2	2.8E-24	[Homo sapiens] [Structural protein] [Extracellular matrix (cuticle and basement membrane); Extracellular (excluding cell wall)] Fibulin 2, an integrin ligand and putative extracellular matrix protein that may participate in supramolecular assemblies and may play a role in hemostatic control
				Talts, J. F. et al. (1995) (supra)
4	7513991CD1	g1165212	1.9E-48	[Homo sapiens] microfibril-associated glycoprotein-2 MAGP-2

Table 2

Polypeptide SEQ ID NO:	Incyte Polypeptide ID	GenBank ID NO: or PROTEOME ID NO:	Probability Score	Annotation
				Gibson, M. A. et al. Further characterization of proteins associated with elastic fiber microfibrils including the molecular cloning of MAGP-2 (MP25). J. Biol. Chem. 271, 1096-1103 (1996).
		344564|MAGP2	1.5E-49	[Homo sapiens] [Structural protein] [Extracellular matrix (cuticle and basement membrane); Extracellular (excluding cell wall)] Microfibril-associated glycoprotein-2, an RGD and cysteine-rich motif-containing glycoprotein of elastic fibril microfibrils in the extracellular matrix; peanut agglutinin binding may be associated with melanoma metastatic potential
				Hatzinikolas, G. et al. The exon structure of the human MAGP-2 gene. Similarity with the MAGP-1 gene is confined to two exons encoding a cysteine-rich region. J. Biol. Chem. 273, 29309-29314 (1998).
		582017|Mfap2	6.2E-19	[Mus musculus] [Structural protein] [Extracellular matrix (cuticle and basement membrane); Extracellular (excluding cell wall)] Microfibrillar-associated protein 2, a component of the elastin-associated microfibrils of the extracellular matrix; terminal galactose and galactosamine modifications of human MFAP2 are associated with the metastatic potential of melanoma cells
				Chen, Y. et al. Structure, chromosomal localization, and expression pattern of the murine Magp gene. J. Biol. Chem. 268, 27381-27389 (1993).
5	7513298CD1	g1381162	3.4E-188	[Homo sapiens] BA46
				Couto, J. R. et al. Cloning and sequence analysis of human breast epithelial antigen BA46 reveals an RGD cell adhesion sequence presented on an epidermal growth factor-like domain. DNA Cell Biol. 15, 281-286 (1996).
		343084|MFGE8	2.6E-189	[Homo sapiens] [Adhesin/agglutinin] [Cytoplasmic; Lipid particles] Milk fat globule-EGF factor 8 protein, a glycoprotein present in breast milk that protects against infection in breast-fed infants, an amyloid precursor with a possible regulatory role in the aorta, a target for breast cancer diagnosis and immunotherapy
				Taylor, M. R. et al. Lactadherin (formerly BA46), a membrane-associated glycoprotein expressed in human milk and breast carcinomas, promotes Arg-Gly-Asp (RGD)-dependent cell adhesion. DNA Cell Biol. 16, 861-869 (1997).

Table 2

Polypeptide SEQ ID NO:	Incyte Polypeptide ID	GenBank ID NO: or PROTEOME ID NO:	Probability Score	Annotation
		582079|Mfge8	1.9E-117	[Mus musculus] [Plasma membrane] Milk fat globule-EGF factor 8 (milk fat globule glycoprotein), contains an epidermal growth factor-like domain with an Arg Gly Asp (RGD) cell adhesion sequence and a domain with similarity to human coagulation factors VIII (human F8) and V (human F5)
				Stubbs, J. D. et al. cDNA cloning of a mouse mammary epithelial cell surface protein reveals the existence of epidermal growth factor-like domains linked to factor VIII-like sequences. Proc. Natl. Acad. Sci. U. S. A. 87, 8417-8421 (1990).
6	7517764CD1	g7634793	3.8E-26	[Homo sapiens] EGF-containing fibulin-like extracellular matrix protein 2
				Katsanis, N. et al. (2000) Isolation of a paralog of the Doyne honeycomb retinal dystrophy gene from the multiple retinopathy critical region on 11q13. Hum. Genet. 106:66-72.
		571186|EFEMP2	2.9E-27	[Homo sapiens] [Structural protein] [Extracellular matrix (cuticle and basement membrane); Basement membrane (extracellular matrix); Extracellular (excluding cell wall)] Egf-containing fibulin-like extracellular matrix protein 2 (Fibulin 4), contains four EGF domains and six calcium-binding EGF domains and may be involved in tumor progression
				Giltay, R. et al. (1999) Sequence, recombinant expression and tissue localization of two novel extracellular matrix proteins, fibulin-3 and fibulin-4. Matrix Biol. 18:469-480.
				Gallagher, W.M. et al. (2001) Human fibulin-4: analysis of its biosynthetic processing and mRNA expression in normal and tumour tissues. FEBS Lett. 489:59-66.
		619022|Efemp2	9.4E-24	[Mus musculus] Egf-containing fibulin-like extracellular matrix protein 2 (Fibulin 4), contains four EGF domains and six calcium-binding EGF domains, binds preferentially to a mutant form of p53 (Trp53) and may be involved in tumor progression
				Gallagher, W.M. et al. (1999) MBP1: a novel mutant p53-specific protein partner with oncogenic properties. Oncogene 18:3608-3616.
7	7517774CD1	g9446402	8.3E-279	[Homo sapiens] integrin beta-subunit
				Sheppard, D. et al. (1990) Complete amino acid sequence of a novel integrin beta subunit (beta 6) identified in epithelial cells using the polymerase chain reaction. J. Biol. Chem. 265:11502-11507.

Table 2

<table>
<tr><th>Polypeptide SEQ ID NO:</th><th>Incyte Polypeptide ID</th><th>GenBank ID NO: or PROTEOME ID NO:</th><th>Probability Score</th><th>Annotation</th></tr>
<tr><td></td><td></td><td>606202|ITGB6</td><td>6.4E-280</td><td>[Homo sapiens] [Adhesin/agglutinin; Receptor (signaling)] [Plasma membrane] Integrin beta 6, member of a family of cell-surface proteins, binds fibronectin, mediates epithelial cell-matrix interactions in development, wound repair, and neoplasia, regulates lung inflammatory response, receptor for foot and mouth disease virus</td></tr>
<tr><td></td><td></td><td></td><td></td><td>Krissansen, G.W. et al. (1992) Chromosomal locations of the genes coding for the integrin beta 6 and beta 7 subunits. Immunogenetics 35:58-61.</td></tr>
<tr><td></td><td></td><td></td><td></td><td>Jackson, T. et al. (2000) The epithelial integrin alphavbeta6 is a receptor for foot-and-mouth disease virus. J. Virol. 74:4949-4956.</td></tr>
<tr><td></td><td></td><td></td><td></td><td>Agrez, M. et al. (1994) The alpha v beta 6 integrin promotes proliferation of colon carcinoma cells through a unique region of the beta 6 cytoplasmic domain. J. Cell. Biol.</td></tr>
<tr><td></td><td></td><td></td><td></td><td>Breuss, J.M. et al. (1995) Expression of the beta 6 integrin subunit in development, neoplasia and tissue repair suggests a role in epithelial remodeling. J. Cell. Sci. 108:2241-</td></tr>
<tr><td></td><td></td><td>618666|Itgb6</td><td>1.9E-255</td><td>[Mus musculus] Integrin beta 6, member of a family of cell-surface proteins, binds fibronectin, mediates epithelial cell-cell and cell-matrix interactions, regulates lung and skin inflammatory responses and fibrosis, binds and activates latent TGF beta 1 (Tgfb1)</td></tr>
<tr><td></td><td></td><td></td><td></td><td>Munger, J.S. et al. (1999) The integrin alpha v beta 6 binds and activates latent TGF beta 1: a mechanism for regulating pulmonary inflammation and fibrosis. Cell 96:319-328.</td></tr>
<tr><td></td><td></td><td></td><td></td><td>Huang, X. et al. (1998) Expression of the human integrin beta6 subunit in alveolar type II cells and bronchiolar epithelial cells reverses lung inflammation in beta6 knockout mice. Am. J. Respir. Cell. Mol. Biol. 19:636-642.</td></tr>
<tr><td></td><td></td><td></td><td></td><td>Huang, X.Z. et al. (1996) Inactivation of the integrin beta 6 subunit gene reveals a role of epithelial integrins in regulating inflammation in the lung and skin. J. Cell. Biol. 133:921-</td></tr>
<tr><td>8</td><td>7518133CD1</td><td>g192262</td><td>1.8E-45</td><td>[Mus musculus] pro-alpha-1 type I collagen</td></tr>
<tr><td></td><td></td><td></td><td></td><td>French, B.T. et al. (1985) Nucleotide sequence of a cDNA clone for mouse pro alpha 1(I) collagen protein. Gene 39:311-312.</td></tr>
<tr><td></td><td></td><td>782887|
LOC129080</td><td>1.8E-243</td><td>[Homo sapiens] Protein containing two collagen triple helix repeats, has a region of high similarity to a region of collagen type V alpha 2 subunit (mouse Col5a2), which may act in connective tissue maintenance and skeletal development</td></tr>
</table>

Table 2

Polypeptide SEQ ID NO:	Incyte Polypeptide ID	GenBank ID NO: or PROTEOME ID NO:	Probability Score	Annotation
		768994|Emu1	1.6E-196	[Mus musculus] Protein containing two collagen triple helix repeats, which are found in some extracellular proteins, has a region of moderate similarity to a region of collagen type VI alpha 1 (human COL6A1), which may be involved in maintaining muscle fiber integrity
9	7520147CD1	g5639939	2.1E-78	[Homo sapiens] oculoglycan
				Hobby, P. et al., Cloning, modeling, and chromosomal localization for a small leucine-rich repeat proteoglycan (SLRP) family member expressed in human eye, Mol. Vis. 6, 72-78
		569342| OPTC	1.5E-79	[Homo sapiens][Structural protein][Extracellular matrix (cuticle and basement membrane);Extracellular (excluding cell wall)] Opticin, member of the leucine-rich repeat protein family found in the extracellular matrix
				Reardon, A. J. et al., Identification in vitreous and molecular cloning of opticin, a novel member of the family of leucine-rich repeat proteins of the extracellular matrix., J Biol Chem 275, 2123-9 (2000).
		584033| Dspg3	6.4E-42	[Mus musculus] Dermatan sulfate proteoglycan 3, a member of the small leucine rich repeat proteoglycan (SLRP) family, expressed in cartilage and testis
				Kurita, K. et al., Occurrence of PG-Lb, a leucine-rich small chondroitin/dermatan sulphate proteoglycan in mammalian epiphyseal cartilage: molecular cloning and sequence analysis of the mouse cDNA., Biochem J 318, 909-14 (1996).
10	7520276CD1	g12584866	1.6E-117	[Homo sapiens] CYR61 protein
				Leng, E. et al., Organization and expression of the cyr61 gene in normal human fibroblasts, J. Biomed. Sci. 9, 59-67 (2002)
		335962| CYR61	2.0E-118	[Homo sapiens][Extracellular (excluding cell wall);Plasma membrane] Cysteine-rich angiogenic inducer 61, a heparin binding protein involved in cell adhesion, cell migration, angiogenesis, and cell proliferation via integrin receptor signaling pathways, may be involved the progression of breast cancer
				Grzeszkiewicz, T. M. et al., CYR61 stimulates human skin fibroblast migration through Integrin alpha vbeta 5 and enhances mitogenesis through integrin alpha vbeta 3, independent of its carboxyl-terminal domain., J Biol Chem 276, 21943-50. (2001).

Table 2

Polypeptide SEQ ID NO:	Incyte Polypeptide ID	GenBank ID NO: or PROTEOME ID NO:	Probability Score	Annotation
		584619| Cyr61	1.4E-108	[Mus musculus][Small molecule-binding protein][Extracellular matrix (cuticle and basement membrane);Extracellular (excluding cell wall);Plasma membrane] Cysteine rich protein 61, an integrin ligand involved in cell adhesion, cell migration, angiogenesis, and cell proliferation, may also play a role in hemostasis; human CYYR61 may be involved the progression of breast cancer
				Chen, C. C. et al., The angiogenic factors cyr61 and connective tissue growth factor induce adhesive signaling in primary human skin fibroblasts., J Biol Chem 276, 10443-52. (2001).
11	7520808CD1	g561659	1.0E-149	[Homo sapiens] receptor of advanced glycosylation end products of proteins
				Sugaya, K. et al., Three genes in the human MHC class III region near the junction with the class II: gene for receptor of advanced glycosylation end products, PBX2 homeobox gene and a notch homolog, human counterpart of mouse mammary tumor gene int-3, Genomics 23, 408-419 (1994)
		618440| AGER	7.4E-151	[Homo sapiens][Receptor (signaling)][Plasma membrane] Receptor for advanced glycation end products, member of the immunoglobulin superfamily that serves as a receptor for advanced glycation end products; high levels of RAGE may be associated with Alzheimer's disease and systemic amyloidosis
				Hofmann, M. A. et al., RAGE mediates a novel proinflammatory axis: a central cell surface receptor for S100/calgranulin polypeptides., Cell 97, 889-901 (1999).
		756758| Ager	4.6E-119	[Rattus norvegicus][Receptor (signaling)] Member of the immunoglobulin superfamily that functions as a receptor for advanced glycation end products
				Hori, O. et al., The receptor for advanced glycation end products (RAGE) is a cellular binding site for amphoterin. Mediation of neurite outgrowth and co-expression of rage and amphoterin in the developing nervous system., J Biol Chem 270, 25752-61 (1995).
				Lander, H. M. et al., Activation of the receptor for advanced glycation end products triggers a p21(ras)-dependent mitogen-activated protein kinase pathway regulated by oxidant stress., J Biol Chem 272, 17810-4 (1997).
12	7520821CD1	g642032	1.4E-24	[Homo sapiens] microfibril-associated glycoprotein

Table 2

Polypeptide SEQ ID NO:	Incyte Polypeptide ID	GenBank ID NO: or PROTEOME ID NO:	Probability Score	Annotation
				Faraco, J. et al., Characterization of the human gene for microfibril-associated glycoprotein (MFAP2), assignment to chromosome 1p36.1-p35, and linkage to D1S170, Genomics 25, 630-637 (1995)
		606230| MFAP2	1.0E-25	[Homo sapiens][Structural protein][Extracellular matrix (cuticle and basement membrane);Extracellular (excluding cell wall)] Microfibrillar-associated protein 2, a component of the elastin-associated microfibrils of the extracellular matrix; terminal galactose and galactosamine modifications are associated with the metastatic potential of
				Segade, F. et al., Identification of a matrix-binding domain in MAGP1 and MAGP2 and intracellular localization of alternative splice forms., J Biol Chem 277, 11050-7. (2002).
		582017| Mfap2	2.3E-21	[Mus musculus][Structural protein][Extracellular matrix (cuticle and basement membrane);Extracellular (excluding cell wall)] Microfibrillar-associated protein 2, a component of the elastin-associated microfibrils of the extracellular matrix; terminal galactose and galactosamine modifications of human MFAP2 are associated with the metastatic potential of melanoma cells
				Segade, F. et al., (supra)
13	7520839CD1	g178529	1.0E-67	[Homo sapiens] amelogenin
				Salido, E. C. et al., The human enamel protein gene amelogenin is expressed from both the X and the Y chromosomes, Am. J. Hum. Genet. 50, 303-316 (1992)
		343950| AMELX	7.8E-69	[Homo sapiens][Structural protein][Extracellular matrix (cuticle and basement membrane);Extracellular (excluding cell wall)] Amelogenin X, a putative extracellular matrix protein of the dental enamel involved in tooth development; mutations in the corresponding gene are associated with amelogenesis imperfecta
				Ravindranath, R. M. et al., The enamel protein amelogenin binds to the N-acetyl-D-glucosamine-mimicking peptide motif of cytokeratins, J Biol Chem 275, 39654-61 (2000).
		579691|Amel	4.7E-62	[Mus musculus][Structural protein][Extracellular matrix (cuticle and basement membrane);Extracellular (excluding cell wall)] Amelogenin, an extracellular matrix protein of the dental enamel involved in mineralization during tooth development; lack of amelogenin causes amelogenesis imperfecta
				Ravindranath, R. M. et al., (supra)

Table 2

Polypeptide SEQ ID NO:	Incyte Polypeptide ID	GenBank ID NO: or PROTEOME ID NO:	Probability Score	Annotation
14	7520891CD1	g178529	3.6E-97	[Homo sapiens] amelogenin
				Salido, E. C. et al., (supra)
		343950|AMELX	2.7E-98	[Homo sapiens][Structural protein][Extracellular matrix (cuticle and basement membrane);Extracellular (excluding cell wall)] Amelogenin X, a putative extracellular matrix protein of the dental enamel involved in tooth development; mutations in the corresponding gene are associated with amelogenesis imperfecta
				Ravindranath, R. M. et al., (supra)
		579691|Amel	5.1E-65	[Mus musculus][Structural protein][Extracellular matrix (cuticle and basement membrane);Extracellular (excluding cell wall)] Amelogenin, an extracellular matrix protein of the dental enamel involved in mineralization during tooth development; lack of amelogenin causes amelogenesis imperfecta
				Ravindranath, R. M. et al., (supra)
15	7514645CD1	g20975688	1.8E-38	[Homo sapiens] (AJ487519) leucine-rich glioma inactivated protein 4
		611258|Lgi1	9.6E-13	[Mus musculus] Leucine-rich glioma inactivated 1, may inhibit cell proliferation; mutations of human LGI1 cause autosomal dominant partial epilepsy with auditory features, and human LGI1 gene rearrangement and low expression are associated with malignant glioma
				Kalachikov, S. et al., Mutations in LGI1 cause autosomal-dominant partial epilepsy with auditory features., Nat Genet 30, 335-41. (2002).
		341690|LGI1	9.6E-13	[Homo sapiens] Leucine-rich glioma inactivated 1, may have a role in neurogenesis and tumor suppression; mutations of the gene cause autosomal dominant partial epilepsy with auditory features, gene rearrangement and low expression are associated with malignant
				Kalachikov, S. et al., (supra)
16	7517776CD1	g200057	1.5E-222	[Mus musculus] neuronal glycoprotein
				Connelly, M. A. et al., PANG, a gene encoding a neuronal glycoprotein, is ectopically activated by intracisternal A-type particle long terminal repeats in murine plasmacytomas., Proc Natl Acad Sci U S A 91, 1337-41 (1994).
		612414|CNTN3	1.1E-241	[Homo sapiens] Protein with strong similarity to plasmacytoma-associated neuronal glycoprotein (rat Pang), which is a neuronal adhesion molecule, contains five immunoglobulin (Ig) domains and four fibronectin type III domains

Table 2

Polypeptide SEQ ID NO:	Incyte Polypeptide ID	GenBank ID NO: or PROTEOME ID NO:	Probability Score	Annotation
				Connelly, M. A. et al., (supra)
		609799| Cntn3	9.8E-220	[Rattus norvegicus][Adhesin/agglutinin;Receptor (signaling)][Unspecified membrane] Plasmacytoma-associated neuronal glycoprotein, neuronal adhesion molecule that may play a role in development of neural networks
				Yoshihara, Y. et al., BIG-1: a new TAG-1/F3-related member of the immunoglobulin superfamily with neurite outgrowth-promoting activity., Neuron 13, 415-26 (1994).
17	7517783CD1	g200057	0.0	[Mus musculus] neuronal glycoprotein
				Connelly, M. A. et al., (supra)
		609799| Cntn3	0.0	[Rattus norvegicus][Adhesin/agglutinin;Receptor (signaling)][Unspecified membrane] Plasmacytoma-associated neuronal glycoprotein, neuronal adhesion molecule that may play a role in development of neural networks
				Yoshihara, Y. et al., (supra)
		612414| CNTN3	0.0	[Homo sapiens] Protein with strong similarity to plasmacytoma-associated neuronal glycoprotein (rat Pang), which is a neuronal adhesion molecule, contains five immunoglobulin (Ig) domains and four fibronectin type III domains
				Connelly, M. A. et al., (supra)
18	7522607CD1	g4519541	5.0E-77	[Mus musculus] thrombospondin type 1 domain
		368602| Mm.42202	3.7E-78	[Mus musculus] Protein containing a type 1 thrombospondin domain, has high similarity to uncharacterized human FLJ14440
		731643| FLJ14440	4.2E-38	[Homo sapiens] Protein containing a type 1 thrombospondin domain, has high similarity to uncharacterized mouse Mm.42202
19	7521142CD1	g13991915	7.6E-58	[Homo sapiens] chondroadherin
				Mansson, B. et al., Association of chondroadherin with collagen type II, J. Biol. Chem. 276, 32883-32888 (2001)
		339770| CHAD	1.2E-58	[Homo sapiens][Adhesin/agglutinin][Extracellular matrix (cuticle and basement membrane);Basement membrane (extracellular matrix);Extracellular (excluding cell wall)] Chondroadherin, a member of the leucine-rich repeat (LRR) family, functions as a cell adhesion molecule of the cartilage extracellular matrix

Table 2

Polypeptide SEQ ID NO:	Incyte Polypeptide ID	GenBank ID NO: or PROTEOME ID NO:	Probability Score	Annotation
				Camper, L. et al., Integrin alpha2beta1 is a receptor for the cartilage matrix protein chondroadherin., J Cell Biol 138, 1159-67. (1997).
				Mansson, B. et al., Association of chondroadherin with collagen type II., J Biol Chem 276, 32883-8. (2001).
		583783|Chad	2.4E-51	[Mus musculus][Adhesin/agglutinin][Extracellular matrix (cuticle and basement membrane);Extracellular (excluding cell wall)] Chondroadherin, a member of the leucine-rich repeat (LRR) family and a putative cell adhesion molecule of the cartilage extracellular
				Landgren, C. et al., The mouse chondroadherin gene: characterization and chromosomal localization., Genomics 47, 84-91 (1998).
				Shen, Z. et al., Chondroadherin expression changes in skeletal development., Biochem J 330, 549-57 (1998). (supra)
20	7521689CD1	g786119	2.0E-48	[Homo sapiens] extracellular matrix protein
				Abrams, W. R. et al., Molecular cloning of the microfibrillar protein MFAP3 and assignment of the gene to human chromosome 5q32-q33.2, Genomics 26, 47-54 (1995)
		344584|MFAP3	1.5E-49	[Homo sapiens][Structural protein][Extracellular matrix (cuticle and basement membrane);Extracellular (excluding cell wall)] Microfibrillar associated protein 3, a component of the elastin-associated microfibrils of the extracellular matrix
				Abrams, W. R. et al., Molecular cloning of the microfibrillar protein MFAP3 and assignment of the gene to human chromosome 5q32-q33.2., Genomics 26, 47-54 (1995).
21	2878775CD1	g13876380	0.0	[Homo sapiens] protocadherin 10
				Wu, Q. et al., A striking organization of a large family of human neural cadherin-like cell adhesion genes, Cell 97, 779-790 (1999)
		743166| PCDH10	0.0	[Homo sapiens] Protocadherin 10, a member of a subclass of the cadherin family of calcium-dependent cell-cell adhesion molecules, contains five cadherin domains and a conserved 17 amino acid cytoplasmic motif, plays a likely role in cell-cell adhesion
				Wolverton, T. et al., Identification and characterization of three members of a novel subclass of protocadherins., Genomics 76, 66-72. (2001).

Table 2

Polypeptide SEQ ID NO:	Incyte Polypeptide ID	GenBank ID NO: or PROTEOME ID NO:	Probability Score	Annotation
		586805| Pcdh10	0.0	[Mus musculus][Plasma membrane] Protocadherin 10, a homophilic cell adhesion molecule that may be involved in the formation of the neural network by segregation of the brain nuclei and mediation of the axonal connections
				Hirano, S. et al., Expression of a novel protocadherin, OL-protocadherin, in a subset of functional systems of the developing mouse brain., J Neurosci 19, 995-1005 (1999).
22	7521207CD1	g16903156	9.8E-26	[Homo sapiens] amelogenin
				Hart, P. S. et al., A nomenclature for X-linked amelogenesis imperfecta, Arch. Oral Biol. 47, 255-260 (2002)
		343950| AMELX	1.2E-23	[Homo sapiens][Structural protein][Extracellular matrix (cuticle and basement membrane);Extracellular (excluding cell wall)] Amelogenin X, a putative extracellular matrix protein of the dental enamel involved in tooth development; mutations in the corresponding gene are associated with amelogenesis imperfecta
				Ravindranath, R. M. et al., The enamel protein amelogenin binds to the N-acetyl-D-glucosamine-mimicking peptide motif of cytokeratins, J Biol Chem 275, 39654-61 (2000).
		579691|Amel	5.5E-20	[Mus musculus][Structural protein][Extracellular matrix (cuticle and basement membrane);Extracellular (excluding cell wall)] Amelogenin, an extracellular matrix protein of the dental enamel involved in mineralization during tooth development; lack of amelogenin causes amelogenesis imperfecta
				Ravindranath, R. M. et al., The enamel protein amelogenin binds to the N-acetyl-D-glucosamine-mimicking peptide motif of cytokeratins, J Biol Chem 275, 39654-61 (2000).
23	7521283CD1	g521108	1.3E-32	[Rattus norvegicus] leucine-rich amelogenin peptide precursor
				Bonass, W. A. et al., Isolation and characterisation of an alternatively-spliced rat amelogenin cDNA: LRAP--a highly conserved, functional alternatively-spliced amelogenin?, Biochim. Biophys. Acta 1219, 690-692 (1994)
		334154| AMELY	1.8E-23	[Homo sapiens][Structural protein][Extracellular matrix (cuticle and basement membrane);Extracellular (excluding cell wall)] Amelogenin, an extracellular matrix protein of the dental enamel, may regulate the formation of crystallites during the secretory stage of enamel development

Table 2

Polypeptide SEQ ID NO:	Incyte Polypeptide ID	GenBank ID NO: or PROTEOME ID NO:	Probability Score	Annotation
				Salido, E. C. et al., The human enamel protein gene amelogenin is expressed from both the X and the Y chromosomes, Am J Hum Genet 50, 303-16 (1992). (supra)
		579691|Amel	4.2E-22	[Mus musculus][Structural protein][Extracellular matrix (cuticle and basement membrane);Extracellular (excluding cell wall)] Amelogenin, an extracellular matrix protein of the dental enamel involved in mineralization during tooth development; lack of amelogenin causes amelogenesis imperfecta
				Lyngstadaas, S. P. et al., A synthetic, chemically modified ribozyme eliminates amelogenin, the major translation product in developing mouse enamel in vivo., Embo Journal 14, 5224-9 (1995). (supra)
24	7522210CD1	g4878035	3.1E-13	[Gallus gallus] neurocan core protein precursor
				Li, H. et al., Coordinate regulation of cadherin and integrin function by the chondroitin sulfate proteoglycan neurocan, J. Cell Biol. 149, 1275-1288 (2000)
		583891|Cspg3	3.9E-13	[Mus musculus][Plasma membrane] Chondroitin sulfate proteoglycan 3 (neurocan), a member of the chondroitin sulfate proteoglycan family, may play a role in modulating neuronal cell adhesion and migration during neurogenesis in the brain
				Oleszewski, M. et al., Characterization of the L1-neurocan-binding site. IMPLICATIONS FOR L1-L1 HOMOPHILIC BINDING, J Biol Chem 275, 34478-85 (2000).
				Oleszewski, M. et al., Integrin and neurocan binding to L1 involves distinct Ig domains., J Biol Chem 274, 24602-10. (1999).
		711568|Cspg3	5.1E-12	[Rattus norvegicus][Small molecule-binding protein][Extracellular matrix (cuticle and basement membrane);Nuclear] Chondroitin sulfate proteoglycan 3 (neurocan), binds hyaluronan, involved in cell adhesion, may play a role in axon guidance, cell interactions, and neurite outgrowth during neurogenesis in the brain
				Friedlander, D. R. et al., The neuronal chondroitin sulfate proteoglycan neurocan binds to the neural cell adhesion molecules Ng-CAM/L1/NILE and N-CAM, and inhibits neuronal adhesion and neurite outgrowth., J Cell Biol 125, 669-80 (1994).
25	7519488CD1	g7239360	1.0E-35	[Homo sapiens] acetylcholinesterase collagen-like tail subunit isoform IV

Table 2

Polypeptide SEQ ID NO:	Incyte Polypeptide ID	GenBank ID NO: or PROTEOME ID NO:	Probability Score	Annotation
				Ohno, K. et al., Human endplate acetylcholinesterase deficiency caused by mutations in the collagen-like tail subunit (ColQ) of the asymmetric enzyme, Proc. Natl. Acad. Sci. U.S.A. 95, 9654-9659 (1998)
				Ohno, K. et al., The spectrum of mutations causing end-plate acetylcholinesterase deficiency, Ann. Neurol. 47, 162-170 (2000)
		342414|COLQ	7.8E-37	[Homo sapiens][Anchor Protein][Extracellular matrix (cuticle and basement membrane); Basement membrane (extracellular matrix); Extracellular (excluding cell wall)] Collagen-like tail subunit of asymmetric acetylcholinesterase (type Q collagen), may attach multimeric complexes of aceylcholinesterase to the basal lamina at the neuromuscular junction; mutation of the corresponding gene causes myasthenic conditions
				Ohno, K. et al., Human endplate acetylcholinesterase deficiency caused by mutations in the collagen-like tail subunit (ColQ) of the asymmetric enzyme, Proc Natl Acad Sci U S A 95, 9654-9 (1998).
				Chitlaru, T. et al., Effect of human acetylcholinesterase subunit assembly on its circulatory residence, Biochem J 354, 613-25. (2001).
		609629|Colq	1.5E-33	[Rattus norvegicus][Structural protein][Extracellular matrix (cuticle and basement membrane); Extracellular (excluding cell wall)] Collagen-like tail subunit of asymmetric acetylcholinesterase (type Q collagen), may attach multimeric complexes of aceylcholinesterase to the basal lamina at the neuromuscular junction; mutation of human COLQ causes myasthenic conditions
				Krejci, E. et al., The mammalian gene of acetylcholinesterase-associated collagen, J Biol Chem 272, 22840-7 (1997).
				Krejci, E. et al., Differences in expression of acetylcholinesterase and collagen Q control the distribution and oligomerization of the collagen-tailed forms in fast and slow muscles, J Neurosci 19, 10672-9 (1999).
26	7519965CD1	g7239363	3.2E-153	[Homo sapiens] acetylcholinesterase collagen-like tail subunit isoform VII
				Ohno, K. et al., Human endplate acetylcholinesterase deficiency caused by mutations in the collagen-like tail subunit (ColQ) of the asymmetric enzyme, Proc. Natl. Acad. Sci. U.S.A. 95, 9654-9659 (1998)

Table 2

Polypeptide SEQ ID NO:	Incyte Polypeptide ID	GenBank ID NO: or PROTEOME ID NO:	Probability Score	Annotation
				Ohno, K. et al., The spectrum of mutations causing end-plate acetylcholinesterase deficiency, Ann. Neurol. 47, 162-170 (2000)
		342414|COLQ	5.4E-148	[Homo sapiens][Anchor Protein][Extracellular matrix (cuticle and basement membrane); Basement membrane (extracellular matrix); Extracellular (excluding cell wall)] Collagen-like tail subunit of asymmetric acetylcholinesterase (type Q collagen), may attach multimeric complexes of aceylcholinesterase to the basal lamina at the neuromuscular junction; mutation of the corresponding gene causes myasthenic conditions
				Ohno, K. et al., Human endplate acetylcholinesterase deficiency caused by mutations in the collagen-like tail subunit (ColQ) of the asymmetric enzyme, Proc Natl Acad Sci U S A 95, 9654-9 (1998).
				Chitlaru, T. et al., Effect of human acetylcholinesterase subunit assembly on its circulatory residence, Biochem J 354, 613-25. (2001).
		420120|Colq	1.3E-116	[Mus musculus][Structural protein; Small molecule-binding protein][Extracellular matrix (cuticle and basement membrane); Extracellular (excluding cell wall)] Collagen-like tail subunit of asymmetric acetylcholinesterase (type Q collagen), may attach multimeric complexes of aceylcholinesterase to the basal lamina at the neuromuscular junction; mutation of human COLQ causes myasthenic conditions
				Feng, G. et al., Genetic analysis of collagen Q: roles in acetylcholinesterase and butyrylcholinesterase assembly and in synaptic structure and function, J Cell Biol 144, 1349-60 (1999).
				Arikawa-Hirasawa, E. et al., Absence of acetylcholinesterase at the neuromuscular junctions of perlecan-null mice, Nat Neurosci 5, 119-23 (2002).
27	7519985CD1	g3298345	7.6E-33	[Homo sapiens] uroplkin 3
				Yuasa, T. et al., Expression of uroplakin Ib and uroplakin III genes in tissues and peripheral blood of patients with transitional cell carcinoma, Jpn. J. Cancer Res. 89, 1-4 (1998)
		568106|UPK3	5.7E-34	[Homo sapiens][Unspecified membrane] Uroplakin 3, a urothelium-specific transmembrane protein, presence in peripheral blood may be a diagnostic marker for metastatic transitional cell carcinoma, may play a role in primary vesicoureteral reflux

Table 2

Polypeptide SEQ ID NO:	Incyte Polypeptide ID	GenBank ID NO: or PROTEOME ID NO:	Probability Score	Annotation
				Yuasa, T. et al., Expression of uroplakin Ib and uroplakin III genes in tissues and peripheral blood of patients with transitional cell carcinoma, Jpn J Cancer Res 89, 879-82 (1998).
				Lobban, E. D. et al., Uroplakin gene expression by normal and neoplastic human urothelium, Am J Pathol 153, 1957-67 (1998).
		746537|Upk3	6.9E-29	[Mus musculus][Plasma membrane] Uroplakin 3, an integral membrane protein required for the formation of normal urothelial plaques in the urothelium, may contribute to the permeability barrier of the urinary tract, plays a role in primary vesicoureteral reflux
				Hu, P. et al., Ablation of uroplakin III gene results in small urothelial plaques, urothelial leakage, and vesicoureteral reflux, J Cell Biol 151, 961-72 (2000).
28	7520002CD1	g181519	6.4E-36	[Homo sapiens] decorin
				Vetter, U. et al., Human decorin gene: intron-exon junctions and chromosomal localization, Genomics 15, 161-168 (1993)
		788335|DCN	4.8E-37	[Homo sapiens][Extracellular matrix (cuticle and basement membrane)] Decorin, a dermatan/chondroitin sulfate proteoglycan that binds to collagen and transforming growth factor beta, negatively controls cell growth and may have a role during organogenesis, deficiency is associated with Marfan syndrome
				Hildebrand, A. et al., Interaction of the small interstitial proteoglycans biglycan, decorin and fibromodulin with transforming growth factor beta, Biochem J 302, 527-34 (1994).
				Shirk, R. A. et al., Altered dermatan sulfate structure and reduced heparin cofactor II-stimulating activity of biglycan and decorin from human atherosclerotic plaque, J Biol Chem 275, 18085-92 (2000).
				Keene, D. R. et al., Decorin binds near the C terminus of type I collagen, J Biol Chem 275, 21801-4 (2000).
		583957|Dcn	7.4E-25	[Mus musculus][Extracellular matrix (cuticle and basement membrane); Extracellular (excluding cell wall)] Decorin, a dermatan/chondroitin sulfate proteoglycan that negatively controls cell growth and may have a role during ontogenesis and organogenesis; deficiency of human DCN is associated with Marfan syndrome
				Danielson, K. G. et al., Targeted disruption of decorin leads to abnormal collagen fibril morphology and skin fragility, J Cell Biol 136, 729-43 (1997).

Table 2

Polypeptide SEQ ID NO:	Incyte Polypeptide ID	GenBank ID NO: or PROTEOME ID NO:	Probability Score	Annotation
				Iozzo, R. V. et al., Cooperative action of germ-line mutations in decorin and p53 accelerates lymphoma tumorigenesis, Proc Natl Acad Sci U S A 96, 3092-7 (1999).
				Santra, M. et al., An anti-oncogenic role for decorin. DOWN-REGULATION OF ErbB2 LEADS TO GROWTH SUPPRESSION AND CYTODIFFERENTIATION OF MAMMARY CARCINOMA CELLS, J Biol Chem 275, 35153-61 (2000).
29	7520014CD1	g5639939	1.3E-90	[Homo sapiens] oculoglycan
				Hobby, P. et al., Cloning, modeling, and chromosomal localization for a small leucine-rich repeat proteoglycan (SLRP) family member expressed in human eye, Mol. Vis. 6, 72-78
		569342|OPTC	1.0E-91	[Homo sapiens][Structural protein][Extracellular matrix (cuticle and basement membrane); Extracellular (excluding cell wall)] Opticin, member of the leucine-rich repeat protein family found in the extracellular matrix
				Reardon, A. J. et al., Identification in vitreous and molecular cloning of opticin, a novel member of the family of leucine-rich repeat proteins of the extracellular matrix, J Biol Chem 275, 2123-9 (2000).
				Hobby, P. et al., Cloning, modeling, and chromosomal localization for a small leucine-rich repeat proteoglycan (SLRP) family member expressed in human eye, Mol Vis 6, 72-8
		582331|Ogn	1.7E-16	[Mus musculus][Structural protein] Osteoglycin, a member of the keratan sulfate proteoglycan group of the small leucine-rich proteoglycan family, may play a role in regulating corneal transparency
				Liu, C. Y. et al., The cloning of mouse keratocan cDNA and genomic DNA and the characterization of its expression during eye development, J Biol Chem 273, 22584-8
30	7520039CD1	g386792	9.5E-58	[Homo sapiens] intercellular adhesion molecule 2
				Dustin ML. et al., Structure and regulation of the leukocyte adhesion receptor LFA-1 and its counterreceptors, ICAM-1 and ICAM-2. Cold Spring Harb Symp Quant Biol 54, Pt 2:753-65 (1989)
		335916|ICAM2	2.4E-58	[Homo sapiens][Adhesin/agglutinin; Ligand][Plasma membrane] Intercellular adhesion molecule 2, a surface glycoprotein and member of the immunoglobulin superfamily, binds the integrin LFA-1 (ITGB2) and promotes cell adhesion during immunological and

Table 2

Polypeptide SEQ ID NO:	Incyte Polypeptide ID	GenBank ID NO: or PROTEOME ID NO:	Probability Score	Annotation
				McLaughlin, F. et al., Tumor necrosis factor (TNF)-alpha and interleukin (IL)-1beta down-regulate intercellular adhesion molecule (ICAM)-2 expression on the endothelium, Cell Adhes Commun 6, 381-400 (1998).
				Hauser, I. A. et al., Differential induction of VCAM-1 on human iliac venous and arterial endothelial cells and its role in adhesion, J Immunol 151, 5172-85 (1993).
				Griffioen, A. W. et al., Endothelial intercellular adhesion molecule-1 expression is suppressed in human malignancies: the role of angiogenic factors, Cancer Res 56, 1111-17
		585029|Icam2	2.2E-50	[Mus musculus][Adhesin/agglutinin; Ligand][Plasma membrane] Intercellular adhesion molecule 2, a surface glycoprotein and member of the immunoglobulin superfamily, binds the integrin LFA-1 (Itgb2) and promotes cell adhesion during immunological and
				Xu, H. et al., Isolation, characterization, and expression of mouse ICAM-2 complementary and genomic DNA, J Immunol 149, 2650-5 (1992).
				Fabry, Z. et al., Adhesion molecules on murine brain microvascular endothelial cells: expression and regulation of ICAM-1 and Lgp 55, J Neuroimmunol 36, 1-11 (1992).
				Gerwin, N. et al., Prolonged eosinophil accumulation in allergic lung interstitium of ICAM-2 deficient mice results in extended hyperresponsiveness, Immunity 10, 9-19 (1999).
31	7520053CD1	g3298345	3.3E-32	[Homo sapiens] uroplkin 3
				Yuasa, T. et al., Expression of uroplakin Ib and uroplakin III genes in tissues and peripheral blood of patients with transitional cell carcinoma, Jpn. J. Cancer Res. 89, 1-4 (1998)
		568106|UPK3	2.5E-33	[Homo sapiens][Unspecified membrane] Uroplakin 3, a urothelium-specific transmembrane protein, presence in peripheral blood may be a diagnostic marker for metastatic transitional cell carcinoma, may play a role in primary vesicoureteral reflux
				Yuasa, T. et al., Expression of uroplakin Ib and uroplakin III genes in tissues and peripheral blood of patients with transitional cell carcinoma, Jpn J Cancer Res 89, 879-82 (1998).
				Lobban, E. D. et al., Uroplakin gene expression by normal and neoplastic human urothelium, Am J Pathol 153, 1957-67 (1998).
		746537|Upk3	3.8E-28	[Mus musculus][Plasma membrane] Uroplakin 3, an integral membrane protein required for the formation of normal urothelial plaques in the urothelium, may contribute to the permeability barrier of the urinary tract, plays a role in primary vesicoureteral reflux

Table 2

Polypeptide SEQ ID NO:	Incyte Polypeptide ID	GenBank ID NO: or PROTEOME ID NO:	Probability Score	Annotation
				Hu, P. et al., Ablation of uroplakin III gene results in small urothelial plaques, urothelial leakage, and vesicoureteral reflux, J Cell Biol 151, 961-72 (2000).
32	7523262CD1	g12652603	7.7E-113	[Homo sapiens] Similar to intercellular adhesion molecule 4, Landsteiner-Wiener blood
		661210|ICAM4	2.3E-69	[Homo sapiens][Adhesin/agglutinin;Ligand][Plasma membrane] Intercellular adhesion molecule 4, an erythrocyte member of the intracellular adhesion molecule (ICAM) family, binds leukocyte-specific integrins, may be involved in erythropoiesis
				Sistonen, P., Linkage of the LW blood group locus with the complement C3 and Lutheran blood group loci., Ann Hum Genet , 239-42 (1984).
		583301|Icam5	1.1E-15	[Mus musculus][Adhesin/agglutinin][Plasma membrane] Intercellular adhesion molecule 5 (telencephalin), an ICAM immunoglobulin family member, induces embryonic neurite outgrowth, may regulate synaptic plasticity; expression of human ICAM5 is decreased in the hippocampus of Alzheimer's patients
				Yoshihara, Y. et al., An ICAM-related neuronal glycoprotein, telencephalin, with brain segment-specific expression., Neuron 12, 541-53 (1994).
33	7523270CD1	g14585875	2.0E-105	[Homo sapiens] proteoglycan link protein
		433054|Crtl1	1.4E-48	[Mus musculus][Structural protein][Extracellular matrix (cuticle and basement membrane);Extracellular (excluding cell wall)] Cartilage linking protein 1, extracellular matrix protein important for the formation of proteoglycan aggregates, and required for normal cartilage and bone formation, may be involved in ovarian follicle development
				Deak, F. et al., Characterization and chromosome location of the mouse link protein gene (Crtl1)., Cytogenet Cell Genet 87, 75-9 (1999).
		609643|Crtl1	1.7E-48	[Rattus norvegicus][Structural protein][Extracellular matrix (cuticle and basement membrane);Extracellular (excluding cell wall)] Cartilage linking protein 1, an extracellular matrix protein that complexes with and stabilizes aggregates of aggrecan and hyaluronan, involved in ovarian follicle development, may be important in brain morphogenesis
				Rhodes, C. et al., Alternative splicing generates two different mRNA species for rat link protein., J Biol Chem 263, 6063-7 (1988).
34	7523287CD1	g13959018	1.5E-75	[Homo sapiens] endothelial cell-selective adhesion molecule

Table 2

Polypeptide SEQ ID NO:	Incyte Polypeptide ID	GenBank ID NO: or PROTEOME ID NO:	Probability Score	Annotation
				Hirata, Ki. et al., Cloning of an immunoglobulin family adhesion molecule selectively expressed by endothelial cells, J. Biol. Chem. 276, 16223-16231 (2001)
35	7521825CD1	g5353510	2.1E-210	[Homo sapiens] emilin precursor
				Doliana, R. et al., EMILIN, a component of the elastic fiber and a new member of the C1q/tumor necrosis factor superfamily of proteins, J. Biol. Chem. 274, 16773-16781 (1999)
		428756|EMILIN	1.5E-211	[Homo sapiens][Structural protein][Extracellular matrix (cuticle and basement membrane);Extracellular (excluding cell wall)] Elastin microfibril interface located protein, an extracellular matrix protein found between amorphous elastin and microfibrils that may play a role in elastin deposition
				Doliana, R. et al., Structure, chromosomal localization, and promoter analysis of the human elastin microfibril interfase located proteIN (EMILIN) gene., J Biol Chem 275, 785-92
		716229| EMILIN-2	3.6E-33	[Homo sapiens] Extracellular glycoprotein EMILIN-2 precursor, a secreted glycoprotein, contains a globular C1q domain, short collagenous stalk, coiled-coil region, proline-rich region, and a cysteine-rich domain (EMI domain), interacts via its gC1q domain with
				Doliana, R. et al., Isolation and characterization of EMILIN-2, a new component of the growing EMILINs family and a member of the EMI domain-containing superfamily., J Biol Chem 276, 12003-11. (2001). (supra)
36	7521844CD1	g17066205	5.6E-131	[Homo sapiens] TSG-6 protein
		432570| TNFAIP6	8.5E-130	[Homo sapiens][Adhesin/agglutinin;Receptor (signalling);Small molecule-binding protein] Tumor necrosis factor induced protein 6, a secreted protein involved in plasmin (PLG) inhibition; found in arthritic synovial fluid and inhibits inflammation; mouse Tnfip6 inhibits inflammation and protects cartilage in arthritis
				Lee, T. H. et al., Transcriptional regulation of TSG6, a tumor necrosis factor- and interleukin-1-inducible primary response gene coding for a secreted hyaluronan-binding protein., J Biol Chem 268, 6154-60 (1993).
		581687|Tnfip6	3.1E-119	[Mus musculus][Structural protein] Tumor necrosis factor induced protein 6, may play a role in matrix formation by the cumulus-oocyte complex, inhibits inflammation and protects arthritic cartilage; human TNFAIP6 is found in arthritic synovial fluid and inhibits

Table 2

Polypeptide SEQ ID NO:	Incyte Polypeptide ID	GenBank ID NO: or PROTEOME ID NO:	Probability Score	Annotation
				Fulop, C. et al., Coding sequence, exon-intron structure and chromosomal localization of murine TNF-stimulated gene 6 that is specifically expressed by expanding cumulus cell-oocyte complexes., Gene 202, 95-102 (1997).
37	7521864CD1	g15196112	1.5E-70	[Homo sapiens] parotid 'o' protein; Po
				Azen, E. A. et al., PRB1, PRB2, and PRB4 coded polymorphisms among human salivary concanavalin-A binding, II-1, and Po proline-rich proteins, Am. J. Hum. Genet. 58, 143-153 (1996)
		337154|PRB4	3.3E-70	[Homo sapiens][Extracellular (excluding cell wall)] Proline-rich protein BstNI subfamily 4, a glycosylated protein that is a member of the salivary proline-rich protein (PRP) gene
				Kim, H. S. et al., The structure and evolution of the human salivary proline-rich protein gene family., Mamm Genome 4, 3-14 (1993).
		341800|PRB1	4.1E-63	[Homo sapiens][Extracellular (excluding cell wall)] Proline-rich protein BstNI subfamily 1 (basic salivary proline rich protein 1), a member of the salivary proline-rich protein (PRP) family, contains tandem repeats that vary in number among gene variants
				Kim, H. S. et al., The structure and evolution of the human salivary proline-rich protein gene family., Mamm Genome 4, 3-14 (1993). (supra)
38	7522020CD1	g2980859	2.7E-103	[Homo sapiens] NKG2C
				Houchins, J. P. et al., DNA sequence analysis of NKG2, a family of related cDNA clones encoding type II integral membrane proteins on human natural killer cells, J. Exp. Med. 173, 1017-1020 (1991)
		336172|KLRC2	5.1E-104	[Homo sapiens][Receptor (signalling)][Plasma membrane] Killer cell lectin-like receptor subfamily C member 2, member of NKG2 family of proteins, contains an extracellular Ca2+ dependent lectin domain, forms natural killing cells activating receptor by association with CD94 (KLRD1) and binds to HLA-E
				Lohwasser, S. et al., Cloning of murine NKG2A, B and C: second family of C-type lectin receptors on murine NK cells., Eur J Immunol 29, 755-61 (1999).

Table 2

Polypeptide SEQ ID NO:	Incyte Polypeptide ID	GenBank ID NO: or PROTEOME ID NO:	Probability Score	Annotation
		568794|KLRC3	2.4E-99	[Homo sapiens][Receptor (signalling)][Plasma membrane] Killer cell lectin-like receptor subfamily C member 3, type II transmembrane molecule with a C-type lectin domain, dimerizes with CD94 and may transmit positive signals, may be invloved in natural killer cell inhibition, activation, and HLA recognition
				Houchins, J. P. et al., DNA sequence analysis of NKG2, a family of related cDNA clones encoding type II integral membrane proteins on human natural killer cells., J Exp Med 173, 1017-20 (1991). (supra)
39	758410CD1	g22652221	1.2E-267	[Mus musculus] (AF410792) alpha 1 type XXIII collagen
		753687|COL5A1	3.9E-95	[Homo sapiens][Structural protein][Extracellular matrix (cuticle and basement membrane);Extracellular (excluding cell wall)] Alpha 1 subunit of type V collagen, involved in skeletal development and contributes to the structure and stability of skin; mutation of corresponding gene causes Ehlers-Danlos syndrome types I and II
				Burrows, N. P. et al., A point mutation in an intronic branch site results in aberrant splicing of COL5A1 and in Ehlers-Danlos syndrome type II in two British families., Am J Hum Genet 63, 390-8 (1998).
		772286| COL11A2	2.0E-91	[Homo sapiens][Structural protein][Extracellular matrix (cuticle and basement membrane);Extracellular (excluding cell wall)] Alpha 2 subunit of type XI collagen, important in hearing and skeletal development; mutation of the corresponding gene causes some forms of Stickler syndrome, oto spondylo megaepiphyseal dysplasia, and non
				Vikkula, M. et al., Autosomal dominant and recessive osteochondrodysplasias associated with the COL11A2 locus, Cell 80, 431-7 (1995).
40	7520759CD1	g35678	2.1E-126	[Homo sapiens] properdin Nolan, K.F. et al. Molecular cloning of the cDNA coding for properdin, a positive regulator of the alternative pathway of human complement. Eur. J. Immunol. 21 (3): 771-
				Nolan, K. F. et al., Characterization of the human properdin gene, Biochem. J. 287 (Pt 1), 291-297 (1992).

Table 2

Polypeptide SEQ ID NO:	Incyte Polypeptide ID	GenBank ID NO: or PROTEOME ID NO:	Probability Score	Annotation
		344704|PFC	1.5E-127	[Homo sapiens][Structural protein][Extracellular (excluding cell wall)] Properdin P factor, a serum protein with a related type-I repeat sequence (TSR), plays a role complement-mediated clearance and inactivation mechanisms of natural and acquired resistance to infection; deficiency leads to fatal bacterial infections
				Goundis, D. et al., Localization of the properdin structural locus to Xp11.23-Xp21.1., Genomics 5, 56-60 (1989).
		323766|Pfc	1.6E-91	[Mus musculus][Structural protein] Properdin factor (complement), may play a role in complement-mediated clearance through the alternative complement pathway; deficiency of human PFC leads to fatal bacterial infections
				Goundis, D. et al., Properdin, the terminal complement components, thrombospondin and the circumsporozoite protein of malaria parasites contain similar sequence motifs., Nature 335, 82-5 (1988).
41	7522915CD1	g156252	1.5E-20	[Caenorhabditis elegans] collagen
				Cox, G. N. et al., Sequence comparisons of developmentally regulated collagen genes of Caenorhabditis elegans, Gene 76, 331-344 (1989).
	7522915CD1	782507|FLJ30681	9.3E-192	[Homo sapiens] Protein containing an epidermal growth factor (EGF)-like domain, has a region of high similarity to a region of collagen type V alpha 2 subunit (human COL5A2), which is an extracellular matrix structural protein involved in connective tissue maintenance
		241095|col-14	1.1E-21	[Caenorhabditis elegans][Structural protein] Collagen
		239340|C15A11.1	4.3E-21	[Caenorhabditis elegans] Protein containing a collagen triple helix repeat, has moderate similarity to a region of alpha 2 subunit of type XI collagen (human COL11A2), which is associated with some forms of Stickler syndrome and Chondrodystrophy with sensorineural deafness
		423291|Col12a1	4.3E-21	[Rattus norvegicus][Structural protein][Extracellular matrix (cuticle and basement membrane);Extracellular (excluding cell wall)] Collagen type XII alpha 1, member of the family of fibril-associated collagens that have interrupted triple helices (FACIT), has increased expression in a model for bladder fibrosis
				Kania, A. M. et al., Structural variation of type XII collagen at its carboxyl-terminal NC1 domain generated by tissue-specific alternative splicing., J Biol Chem 274, 22053-9 (1999).

Table 2

Polypeptide SEQ ID NO:	Incyte Polypeptide ID	GenBank ID NO: or PROTEOME ID NO:	Probability Score	Annotation
42	7522936CD1	g7239359	2.1E-133	[Homo sapiens] acetylcholinesterase collagen-like tail subunit isoform III
				Ohno, K. et al., Human endplate acetylcholinesterase deficiency caused by mutations in the collagen-like tail subunit (ColQ) of the asymmetric enzyme, Proc. Natl. Acad. Sci. U.S.A. 95, 9654-9659 (1998)
42		342414|COLQ	3.3E-93	[Homo sapiens][Anchor Protein][Extracellular matrix (cuticle and basement membrane);Basement membrane (extracellular matrix);Extracellular (excluding cell wall)] Collagen-like tail subunit of asymmetric acetylcholinesterase (type Q collagen), may attach multimeric complexes of aceylcholinesterase to the basal lamina at the neuromuscular junction; mutation of the corresponding gene causes myasthenic conditions
				Ohno, K. et al., The spectrum of mutations causing end-plate acetylcholinesterase deficiency., Ann Neurol 47, 162-70. (2000).
		609629|Colq	2.5E-88	[Rattus norvegicus][Structural protein][Extracellular matrix (cuticle and basement membrane);Extracellular (excluding cell wall)] Collagen-like tail subunit of asymmetric acetylcholinesterase (type Q collagen), may attach multimeric complexes of aceylcholinesterase to the basal lamina at the neuromuscular junction; mutation of human COLQ causes myasthenic conditions
				Krejci, E. et al., The mammalian gene of acetylcholinesterase-associated collagen., J Biol Chem 272, 22840-7 (1997).

Table 3

SEQ ID NO:	Incyte Polypeptide ID	Amino Acid Residues	Signature Sequences, Domains and Motifs	Analytical Methods and Databases
1	7513225CD1	1327	signal_cleavage: M1-A30	SPSCAN
			Signal Peptide: M1-A30, M1-L31	HMMER
			Epidermal growth factor-like domain: E847-I882, E804-I843, P762-V800, T487-L524	HMMER_SMART
			Calcium-binding EGF-like domain: D801-I843, V845-I882, P759-V800	HMMER_SMART
			Low-density lipoprotein-receptor YWTD domain: K1173-R1217, E1130-R1172, R1218-R1260, K1086-E1128, R1261-T1301	HMMER_SMART
			Thyroglobulin type I repeats: I912-P961, V992-V1040	HMMER_SMART
			EGF-like domain: C805-C842, C763-C799, C848-C881, C488-C523	HMMER_PFAM
			Low-density lipoprotein receptor repeats: R1150-I1191, G1193-F1236, R1106-I1148, K1238-D1278	HMMER_PFAM
			Thyroglobulin type-1 repeat: C892-C957, C971-C1036	HMMER_PFAM
			Thyroglobulin type-1 repeat IPB000716: H911-C933, H911-C926, C933-Q947	BLIMPS_BLOCKS
			Calcium-binding EGF-like domain IPB001881: C786-G796, C818-C829	BLIMPS_BLOCKS
			GLYCOPROTEIN EGF-LIKE DOMAIN PRECURSOR BASEMENT MEMBRANE EXTRACELLULAR MATRIX SIGNAL CALCIUM-BINDING PD014162: P525-P762 P955-P968 E526-L687	BLAST_PRODOM
			GLYCOPROTEIN EGF-LIKE DOMAIN NIDOGEN-2 PRECURSOR NID2 OSTEONIDOGEN BASEMENT MEMBRANE EXTRACELLULAR PD132287: A187-G203 E357-T487 S272-E441	BLAST_PRODOM
			GLYCOPROTEIN EGF-LIKE DOMAIN PROTEIN PRECURSOR BASEMENT MEMBRANE EXTRACELLULAR MATRIX SIGNAL PD007677: V77-G285	BLAST_PRODOM
			GLYCOPROTEIN EGF-LIKE DOMAIN NIDOGEN2 PRECURSOR NID2 OSTEONIDOGEN BASEMENT MEMBRANE EXTRACELLULAR PD153979: F1280-K1327	BLAST_PRODOM
			THYROGLOBULIN TYPE I REPEAT DM00280 |P04441|194-255: T890-T948 |P31226|108-174: H911-C957 |P10247|195-256: T890-T948 |P04233|211-272: T890-T948	BLAST_DOMO

Table 3

SEQ ID NO:	Incyte Polypeptide ID	Amino Acid Residues	Signature Sequences, Domains and Motifs	Analytical Methods and Databases
			Potential Phosphorylation Sites: S42 S84 S147 S250 S274 S340 S348 S359 S370 S431 S473 S643 S681 S697 S824 S831 S867 S1157 S1170 S1210 S1275 T245 T304 T550 T626 T773 T780 T890 T964 T969 T985 T1048 T1247 T1253 T1258 T1324	MOTIFS
			Potential Glycosylation Sites: N417 N658 N693 N703 N1076	MOTIFS
			Aspartic acid and asparagine hydroxylation site: C501-C512, C818-C829, C859-C870	MOTIFS
			EGF-like domain signature 2: C786-C799, C827-C842, C868-C881	MOTIFS
			Calcium-binding EGF-like domain pattern signature: D801-C827	MOTIFS
			Thyroglobulin type-1 repeat signature: H911-G940, Y991-G1020	MOTIFS
2	7513288CD1	2110	signal_cleavage: M1-A19	SPSCAN
			Signal Peptide: M1-A19, M1-E21, M1-G22, M1-V24	HMMER
			Fibronectin type 3 domain: L803-S882 L985-S1062 L1796-S1873, L893-S974, V623-P703, M1619-S1697, L1708-S1785, L712-G794, E1529-S1608, T1165-T1245, R1073-R1152, L1256-R1334, L1438-T1517, E1347-K1427	HMMER_SMART
			Epidermal growth factor-like domain: R376-G404 S469-R497 Q500-A528 S531-K559 S593-S621, E189-S217, I251-N280, L283-S311, K407-S435, R562-G590, R438-S466, A220-S248, I314-G342, T345-S373	HMMER_SMART
			Fibrinogen-related domains (FReDs): P1888-S2098	HMMER_SMART
			EGF-like domain: C377-C403 C532-C558 C408-C434, C501-C527, C563-C589, C470-C496, C284-C310, C594-C620, C315-C341, C185-C216, C252-C279, C221-C247, C346-C372, C439-C465	HMMER_PFAM
			Fibrinogen beta and gamma chains, C-terminal globular domain: F1889-S2098	HMMER_PFAM
			Fibronectin type III domain: L1796-S1873, L893-S974, M1619-S1697, L985-S1062, L803-S882, V623-S701, L1708-S1785, L712-L795, T1165-S1243, L1256-S1339, E1529-S1608, A1074-S1157, L1438-T1517, E1347-S1423	HMMER_PFAM
			Cytosolic domain: M1-Q6 Transmembrane domain: L7-L25 Non-cytosolic domain: K26-A2110	TMHMMER
			Fibrinogen beta and gamma chains C-terminal globular domain: IPB002181: M1897-L1911, V1921-G1957, E1962-T1974, Y2008-S2022, N2044-S2073, S2073-P2097	BLIMPS_BLOCKS

Table 3

SEQ ID NO:	Incyte Polypeptide ID	Amino Acid Residues	Signature Sequences, Domains and Motifs	Analytical Methods and Databases
2			Fibrinogen beta and gamma chains C-terminal domain signature: D2035-E2085	PROFILESCAN
			PRECURSOR GLYCOPROTEIN SIGNAL FIBRINOGEN BLOOD COAGULATION CHAIN PLASMA PROTEIN PLATELET PD001241: T306-I314 K1891-S2087 S1873-P2097	BLAST_PRODOM
			GLYCOPROTEIN PRECURSOR TENASCIN SIGNAL MATRIX CYTOTACTIN ANTIGEN CELL TENASCIN X EGF-LIKE PD004440: M4-R137	BLAST_PRODOM
			PROTEIN TRANSCRIPTIONAL REPEAT TRANSCRIPTION REGULATION DNA-BINDING NUCLEAR SHUTTLE CRAFT PUTATIVE PD014613: C161-P625 C172-C604	BLAST_PRODOM
			GLYCOPROTEIN TENASCIN TENASCIN X ANTIGEN PRECURSOR MATRIX CELL X TN HEXABRACHION PD000928: L1353-T1435 L1535-A1618	BLAST_PRODOM
			FIBRINOGEN BETA/GAMMA DM00531 |P24821|1946-2187: P1856-N2099 |S19694|1492-1734: P1856-N2099 |P10039|1554-1796: P1856-N2099 |JH0675|1096-1338: S1857-N2099	BLAST_DOMO
			Potential Phosphorylation Sites: S35 S86 S124 S135 S162 S186 S373 S506 S606 S616 S705 S722 S760 S767 S807 S875 S903 S922 S931 S1125 S1161 S1307 S1384 S1396 S1423 S1475 S1570 S1621 S1638 S1648 S1798 S1850 S2000 S2032 S2048 T40 T101 T152 T212 T262 T306 T337 T399 T430 T492 T523 T585 T726 T799 T800 T801 T847 T852 T867 T891 T947 T983 T987 T1071 T1081 T1095 T1186 T1234 T1277 T1344 T1445 T1471 T1489 T1526 T1550 T1566 T1593 T1706 T1720 T1764 T1773 T1789 T1794 T1803 T1815 T1836 T1861 T2018 Y768 Y1393 Y1946	MOTIFS
			Potential Glycosylation Sites: N38 N166 N184 N327 N788 N1018 N1034 N1079 N1093 N1119 N1184 N1210 N1275 N1301 N1354 N1364 N1394 N1443 N1718 N2071	MOTIFS
			Cell attachment sequence: R877-D879	MOTIFS
			EGF-like domain signature 1: C174-C185, C205-C216, C236-C247, C268-C279, C299-C310, C330-C341, C361-C372, C392-C403, C423-C434, C454-C465, C485-C496, C516-C527, C547-C558, C578-C589, C609-C620	MOTIFS

Table 3

SEQ ID NO:	Incyte Polypeptide ID	Amino Acid Residues	Signature Sequences, Domains and Motifs	Analytical Methods and Databases
			EGF-like domain signature 2: C174-C185, C205-C216, C236-C247, C268-C279, C299-C310, C330-C341, C361-C372, C392-C403, C423-C434, C454-C465, C485-C496, C516-C527, C547-C558, C578-C589, C609-C620	MOTIFS
3	7513607CD1	393	signal_cleavage: M1-A23	SPSCAN
			Signal Peptide: M1-P21, M1-A23, M1-S29, M1-C27, M1-Q25	HMMER
			Domain abundant in complement control protein: C81-C134	HMMER_SMART
			Epidermal growth factor-like domain: E139-Q172, E227-E273	HMMER_SMART
			Calcium-binding EGF-like domain: D224-E273, E139-Q172	HMMER_SMART
			EGF-like domain: C140-C171, C228-C272	HMMER_PFAM
			Sushi domain (SCR repeat): C81-C134	HMMER_PFAM
			Type II EGF-like signature PR00010: G136-N147, G148-V155, N156-R166	BLIMPS_PRINTS
			Type III EGF-like signature PR00011: E153-C171	BLIMPS_PRINTS
			Sushi domain proteins (SCR repeat proteins): D100-F111, G125-C134	BLIMPS_PFAM
			Potential Phosphorylation Sites: S5 S324 S343 S389 T86 T150	MOTIFS
			Potential Glycosylation Sites: N124 N261	MOTIFS
			EGF-like domain signature 1: C160-C171	MOTIFS
			Calcium-binding EGF-like domain pattern signature: D224-C250	MOTIFS
4	7513991CD1	148	signal_cleavage: M1-W21	SPSCAN
			Signal Peptide: M1-S28	HMMER
			MICROFIBRIL-ASSOCIATED GLYCOPROTEIN PRECURSOR EXTRACELLULAR MATRIX SIGNAL MAGP MAGP1 SULFATATION MAGP2 PD013288: T38-Q117	BLAST_PRODOM
			MICROFIBRIL-ASSOCIATED GLYCOPROTEIN PRECURSOR MAGP2 MP25 EXTRACELLULAR MATRIX SIGNAL PD034854: M1-T57	BLAST_PRODOM
			MICROFIBRIL-ASSOCIATED GLYCOPROTEIN PRECURSOR MAGP2 MP25 EXTRACELLULAR MATRIX SIGNAL PD034853: A119-L148	BLAST_PRODOM
			Potential Phosphorylation Sites: S28 S87 T41	MOTIFS
			Cell attachment sequence: R30-D32	MOTIFS
5	7513298CD1	343	signal_cleavage: M1-A23	SPSCAN
			Signal Peptide: M1-A17, M1-S19, M1-L21, M1-A23, M1-C16, M1-E25, M1-P29	HMMER

Table 3

SEQ ID NO:	Incyte Polypeptide ID	Amino Acid Residues	Signature Sequences, Domains and Motifs	Analytical Methods and Databases
			Coagulation factor 5/8 C-terminal domain: E25-C181, G185-C343	HMMER_SMART
			F5/8 type C domain: P189-L340, P29-L178	HMMER_PFAM
			Coagulation factor 5/8 type C domain (FA58C) IPB000421: T260-N265, V269-W286, I299-C343	BLIMPS_BLOCKS
			GLYCOPROTEIN PRECURSOR SIGNAL FACTOR REPEAT PROTEIN NEUROPILIN CELL DOMAIN COAGULATION PD000875: P189-L340 P29-L178	BLAST_PRODOM
			DISCOIDIN I N-TERMINAL DM00516 |P21956|338-462: F216-C343, W57-C181; 177-301: Q55-G180, W218-G342 |A42580|2085-2210: P220-C343, P59-L183 |P12259|2095-2223: W218-C343, H56-L183	BLAST_DOMO
			Potential Phosphorylation Sites: S46 S76 S77 S196 S207 S251 S274 S296 S308 T154 T173	MOTIFS
			Potential Glycosylation Sites: N194 N281 N285 N306	MOTIFS
			Coagulation factors 5/8 type C domain (FA58C) signature 1: A72-G101, A233-G262	MOTIFS
			Coagulation factors 5/8 type C domain (FA58C) signature 2: P165-C181, P327-C343	MOTIFS
6	7517764CD1	110	signal_cleavage: M1-P27	SPSCAN
			Signal Peptide: M1-A25, M1-P27, S6-P27, G10-P27, M1-S30	HMMER
			TNFR/NGFR family cysteine-rich region IPB001368: P3-L13	BLIMPS_BLOCKS
			Potential Phosphorylation Sites: S26 S35	MOTIFS
7	7517774CD1	724	Signal Peptide: M1-G21	HMMER
			Integrin beta subunits (N-terminal portion of extracellular region): T30-C454	HMMER_SMART
			domain found in Plexins, Semaphorins and Integrins: G22-Q71	HMMER_SMART
			Integrins, beta chain: T30-C454	HMMER_PFAM
			Non-cytosolic domain: M1-N643 Transmembrane domain: I644-W666 Cytosolic domain: K667-C724	TMHMMER
			EGF-like domain IPB000561: C499-G507	BLIMPS_BLOCKS

Table 3

SEQ ID NO:	Incyte Polypeptide ID	Amino Acid Residues	Signature Sequences, Domains and Motifs	Analytical Methods and Databases
7			Integrin beta, C-terminus IPB001169: R58-I75, L121-K161, E162-C197, I213-H264, L266-L295, L307-V350, G351-A392, C490-C512, C519-C540, C550-C569	BLIMPS_BLOCKS
			Integrins beta chain cysteine-rich domain signature: G507-S562	PROFILESCAN
			Integrin beta subunit signature PR01186: A28-C44, C59-P78, Q103-G116, Y131-L149, R171-P190, C204-F223, D237-R259, H264-D279, E308-T331, L344-Y368, C527-C540, P642-G659, G659-E677, T694-T704	BLIMPS_PRINTS
			INTEGRIN GLYCOPROTEIN CELL ADHESION TRANSMEMBRANE REPEAT PRECURSOR EXTRACELLULAR SUBUNIT SIGNAL PD001811: T30-C454 PD001794: T549-Y710	BLAST_PRODOM
			INTEGRIN GLYCOPROTEIN CELL ADHESION TRANSMEMBRANE REPEAT SUBUNIT PRECURSOR EXTRACELLULAR MATRIX PD149771: D455-C540	BLAST_PRODOM
			INTEGRINS BETA CHAIN CYSTEINE-RICH DOMAIN DM00846|P18564|1-442: M1-A443 DM00846|P18084|7-451: C23-A443 DM00846|P05106|9-449: C23-A443 DM00846|P12607|2-453: Q20-D442	BLAST_DOMO
			Potential Phosphorylation Sites: S80 S144 S176 S270 S297 S366 S373 S407 S464 S528 S562 S565 S593 S632 S671 T30 T151 T188 T189 T305 T514 T555 T591 T622 T704 T709	MOTIFS
			Potential Glycosylation Sites: N48 N97 N260 N387 N396 N463 N471 N511 N707	MOTIFS
			Cell attachment sequence: R530-D532	MOTIFS
			EGF-like domain signature 1: C479-C490, C499-C510	MOTIFS
			EGF-like domain signature 2: C499-C510	MOTIFS
			Integrins beta chain cysteine-rich domain signature: C527-C540	MOTIFS
8	7518133CD1	445	signal_cleavage: M1-A21	SPSCAN
			Signal Peptide: M1-A21, M1-A22, P4-A22, M1-S24, M1-A28	HMMER
			Collagen triple helix repeat (20 copies): G208-P267, G299-R358	HMMER_PFAM

Table 3

SEQ ID NO:	Incyte Polypeptide ID	Amino Acid Residues	Signature Sequences, Domains and Motifs	Analytical Methods and Databases
			Cytosolic domain: M1-A6 Transmembrane domain: W7-P29 Non-cytosolic domain: F30-G445	TMHMMER
			COLLAGEN ALPHA PRECURSOR CHAIN REPEAT SIGNAL CONNECTIVE TISSUE EXTRACELLULAR MATRIX PD000007: Q164-G238, P165-G256, G199-P277, G223-G332, P291-E369, G293-G368, G293-E369	BLAST_PRODOM
			SIMILAR TO CUTICULAR COLLAGEN PD067228: E163-R270, D191-G305, G217-G314, P224-G317, G235-G332, P257-Q357, P291-G368, P291-K370, P297-P375, P306-Q384	BLAST_PRODOM
			PRECOLLAGEN P PRECURSOR SIGNAL PD072959: G181-P277, G223-G329, G229-G341, G247-P349, G256-P367, G293-H372	BLAST_PRODOM
			FIBRILLAR COLLAGEN CARBOXYL-TERMINAL DM00019|P04258|434-662: E163-G368, P167-G359, G181-L377, P249-P375 DM00019|P12107|1270-1494: Q164-G329, P168-G365, G181-G368, G181-K370, G202-G368, G238-P375, P291-G427 DM00019|S18803|1304-1528: P165-G314, P168-G368, P176-K370, G181-G368, G205-G368, G238-P375, G293-S376 DM00019|P20908|1300-1524: P167-G308, P167-G368, P168-G368, P176-K370, G202-G368, G232-P375, G293-S376	BLAST_DOMO
			Potential Phosphorylation Sites: S31 S72 S103 S188 S440 T80 T89 T122 T146 T160 T172 T242 T282 T395	MOTIFS
			Potential Glycosylation Sites: N51 N138	MOTIFS
9	7520147CD1	279	signal_cleavage: M1-Q15	SPSCAN
			Signal Peptide: M1-Q15, M1-G18, M1-A20, M1-P23, M1-S21, M1-T19	HMMER
			Leucine-rich repeats, typical (most populated) subfamily: A124-L146, M193-S216, L147-G169, L217-D237	HMMER_SMART
			Leucine Rich Repeat: K125-H148, K195-P214, S216-E239, G169-E194, Q247-P271	HMMER_PFAM
			Leucine-rich repeat signature PR00019: L196-I209, M193-L206	BLIMPS_PRINTS

Table 3

SEQ ID NO:	Incyte Polypeptide ID	Amino Acid Residues	Signature Sequences, Domains and Motifs	Analytical Methods and Databases
			PROTEOGLYCAN PRECURSOR SIGNAL GLYCOPROTEIN REPEAT LEUCINE REPEAT CONNECTIVE TISSUE EXTRACELLULAR MATRIX PD006710: L199-L273	BLAST_PRODOM
			LEUCINE-RICH ALPHA-2-GLYCOPROTEIN DM01068|A41781|208-283: L173-R246	BLAST_DOMO
			Potential Phosphorylation Sites: S63 S137 S168 S216 T244	MOTIFS
			Potential Glycosylation Sites: N259	MOTIFS
10	7520276CD1	245	signal_cleavage: M1-S24	SPSCAN
			Signal Peptide: M1-L21, M1-A22, M1-S24, M1-A28	HMMER
			Insulin growth factor-binding protein homologs: S24-A93	HMMER_SMART
			von Willebrand factor (vWF) type C domain: C100-C163	HMMER_SMART
			Insulin-like growth factor binding protein: C26-G97	HMMER_PFAM
			von Willebrand factor type C domain: C100-C163	HMMER_PFAM
			C-terminal cystine knot IPB000359: V46-Q59, C100-P148	BLIMPS_BLOCKS
			Insulin-like growth factor-binding protein IPB000867: S24-L34, D48-D63, Y107-C134	BLIMPS_BLOCKS
			von Willebrand factor, type C repeat IPB001007: C39-G49, L76-C91, C117-G126	BLIMPS_BLOCKS
			Fibronectin type I repeat signature PR00012: C121-G129	BLIMPS_PRINTS
			von Willebrand factor type D domain proteins. PF00094: C121-C130	BLIMPS_PFAM
			Insulin-like growth factor binding proteins signature: A28-T85	PROFILESCAN
			PROTEIN PRECURSOR GROWTH FACTOR BINDING SIGNAL CYR61 CEF10 GIG1 INSULINLIKE PD016694: D166-P211	BLAST_PRODOM
			PRECURSOR SIGNAL REPEAT GLYCOPROTEIN PROTEIN FACTOR GROWTH COLLAGEN EGFLIKE DOMAIN PD000826: C100-C163	BLAST_PRODOM
			INSULIN-LIKE GROWTH FACTOR BINDING PROTEINS DM02941|P19336|4-275: A10-P211 DM02941|P29279|3-250: L9-E165 DM02941|P51609|1-242: L9-D164 DM02941|P48745|12-257: R7-I168	BLAST_DOMO
			Potential Phosphorylation Sites: S2 S167 S188 S206 S217 T122	MOTIFS
			Insulin-like growth factor binding proteins signature: G49-C64	MOTIFS
			VWFC domain signature: C117-C163	MOTIFS

Table 3

SEQ ID NO:	Incyte Polypeptide ID	Amino Acid Residues	Signature Sequences, Domains and Motifs	Analytical Methods and Databases
11	7520808CD1	325	signal_cleavage: M1-A23	SPSCAN
			Signal Peptide: M1-G22, M1-A23, M1-Q24	HMMER
			Immunoglobulin: A23-Y118, P244-Q324	HMMER_SMART
			Cytosolic domain: M1-A6 Transmembrane domain: V7-I26 Non-cytosolic domain: T27-N325	TMHMMER
			Intercellular adhesion molecule/vascular cell adhesion molecule-1 signature PR01472: G246-P262, Y118-S131, I30-P45	BLIMPS_PRINTS
			GLYCOPROTEIN PRECURSOR CELL SI. PD00015: G252-C259, P265-W271	BLIMPS_PRODOM
			GLYCOPROTEIN ANTIGEN PRECURSOR IMMUNOGLOBULIN PD02327: V115-I126, T143-L164, S209-A223	BLIMPS_PRODOM
			PRECURSOR SIGNAL IMMUNOGLOBULIN FOLD GLYCOPROTEIN TRANSMEMBRANE CELL ANTIGEN ADHESION RECEPTOR PD004088: N81-S209	BLAST_PRODOM
			ADVANCED GLYCOSYLATION END PRODUCTSPECIFIC RECEPTOR PRECURSOR FOR PRODUCTS IMMUNOGLOBULIN FOLD PD013100: M1-P80	BLAST_PRODOM
			ADVANCED GLYCOSYLATION END PRODUCTSPECIFIC RECEPTOR PRECURSOR FOR PRODUCTS IMMUNOGLOBULIN FOLD PD150896: M193-W230	BLAST_PRODOM
			IMMUNOGLOBULIN DM00001|I61596|125-228: E125-V229 DM00001|I61596|20-109: V20-K110 DM00001|I61596|230-311: W230-D274	BLAST_DOMO
			Potential Phosphorylation Sites: S129 S172 S290 S322 T27 T55 T177 T316	MOTIFS
			Potential Glycosylation Sites: N25 N81	MOTIFS
12	7520821CD1	58	signal_cleavage: M1-A17	SPSCAN
			Signal Peptide: M1-A17, M1-Q18, M1-Y21, M1-G19	HMMER
			Intercellular adhesion molecule/vascular cell adhesion molecule-1 signature PR01472: G246-P262, Y118-S131, I30-P45, P45-V63	BLIMPS_PRINTS
			MICROFIBRILASSOCIATED GLYCOPROTEIN PRECURSOR EXTRACELLULAR MATRIX SIGNAL MAGP MAGP1 SULFATATION MAGP2 PD013288: M1-Q51	BLAST_PRODOM

Table 3

SEQ ID NO:	Incyte Polypeptide ID	Amino Acid Residues	Signature Sequences, Domains and Motifs	Analytical Methods and Databases
13	7520839CD1	151	signal_cleavage: M1-A16	SPSCAN
			Signal Peptide: M1-A16, M1-P18, M1-T21, M1-P22, M1-W25	HMMER
			Amelogenin: M1-D151	HMMER_PFAM
			Cytosolic domain: Q27-D151 Transmembrane domain: W4-Y26 Non-cytosolic domain: M1-T3	TMHMMER
			Synapsin IPB001359: L79-P96	BLIMPS_BLOCKS
			Osteonectin domain IPB001999: T3-P18	BLIMPS_BLOCKS
			Small proline-rich protein signature PR00021: V92-P104, P111-P120, N78-Y87	BLIMPS_PRINTS
			Vinculin signature PR00806: M109-P119	BLIMPS_PRINTS
			AMELOGENIN ISOFORM PRECURSOR EXTRACELLULAR MATRIX PHOSPHORYLATION ENAMEL REPEAT SIGNAL ALTERNATIVE PD007277: H33-A83, M1-P132, V19-P120	BLAST_PRODOM
			PROTEIN REPEAT SIGNAL PRECURSOR PRION GLYCOPROTEIN NUCLEAR GPIANCHOR BRAIN MAJOR PD001091: P46-E148, Q27-P126	BLAST_PRODOM
			AMELOGENIN DM05282|P45559|128-195: Q84-D151	BLAST_DOMO
			GLIADIN DM00406|P45559|81-127: Q40-P86	BLAST_DOMO
			PROLINE; RICH; PISTIL; EXTENSIN; DM04077|S23737|54-270: P31-P132	BLAST_DOMO
			H-A-P-P REPEAT DM08271|S25299|69-249: Y26-P129	BLAST_DOMO
			Potential Phosphorylation Sites: S28 T142 T145	MOTIFS
14	7520891CD1	175	signal_cleavage: M1-A16	SPSCAN
			Signal Peptide: M1-A16, M1-P18, M1-T21, M1-P22, M1-W25	HMMER
			Amelogenin: M1-D175	HMMER_PFAM
			Cytosolic domain: Q27-D175 Transmembrane domain: W4-Y26 Non-cytosolic domain: M1-T3	TMHMMER
			Synapsin IPB001359: L103-P120	BLIMPS_BLOCKS
			Osteonectin domain IPB001999: T3-P18	BLIMPS_BLOCKS
			Small proline-rich protein signature PR00021: V116-P128, P135-P144, N102-Y111	BLIMPS_PRINTS

Table 3

SEQ ID NO:	Incyte Polypeptide ID	Amino Acid Residues	Signature Sequences, Domains and Motifs	Analytical Methods and Databases
			AMELOGENIN ISOFORM PRECURSOR EXTRACELLULAR MATRIX PHOSPHORYLATION ENAMEL REPEAT SIGNAL ALTERNATIVE PD007277: V19-A107, M1-P144	BLAST_PRODOM
			PROTEIN REPEAT SIGNAL PRECURSOR PRION GLYCOPROTEIN NUCLEAR GPIANCHOR BRAIN MAJOR PD001091: P70-E172	BLAST_PRODOM
			AMELOGENIN DM05282|P45559|128-195: Q108-D175	BLAST_DOMO
			GLIADIN DM00406|P45559|81-127: Q64-P110	BLAST_DOMO
			PROLINE; RICH; PISTIL; EXTENSIN DM04077|S23737|54-270: H57-P156	BLAST_DOMO
			H-A-P-P REPEAT DM08271|S25299|69-249: L20-P156, P31-P153	BLAST_DOMO
			Potential Phosphorylation Sites: S28 T166 T169	MOTIFS
15	7514645CD1	81	signal_cleavage: M1-A19	SPSCAN
			Signal Peptide: M1-V16, M1-A19, M1-R21, M1-P22, M1-G25	HMMER
			Osteonectin domain IPB001999: A4-A19	BLIMPS_BLOCKS
16	7517776CD1	749	Signal Peptide: M2-G19M2-Q24, M2-G18	HMMER
			Fibronectin type 3 domain: P493-G579, E596-G682	HMMER_SMART
			Immunoglobulin: K311-T392, P401-R490, R129-R217, S35-A119	HMMER_SMART
			Immunoglobulin C-2 Type: Q317-G381, S41-G107, S407-D479, R135-T203	HMMER_SMART
			Fibronectin type III domain: P493-S582, P595-S685	HMMER_PFAM
			Immunoglobulin domain: G319-A376, G409-V474, G137-V198, D43-A102	HMMER_PFAM
			I type Ig domains from SCOP: A302-I397, T398-P496	HMMER_INCY
			Ig superfamily from SCOP: F28-A119, I397-P495, F305-T395, K125-P213	HMMER_INCY
			PRECURSOR SIGNAL CONTACTIN CELL ADHESION NEUROFASCIN GLYCOPROTEIN GP135 IMMUNOGLOBULIN FOLD PD001890: L583-P697	BLAST_PRODOM
			NEURAL CELL ADHESION MOLECULE CLOSE HOMOLOGUE OF L1 L1LIKE PROTEIN PD066559: E377-G491	BLAST_PRODOM
			PRECURSOR SIGNAL ADHESION CELL GLYCOPROTEIN IMMUNOGLOBULIN FOLD REPEAT MOLECULE NEURAL PD003129: T103-L216	BLAST_PRODOM
			GLYCOPROTEIN NEURONAL PLASMACYTOMAASSOCIATED PANG BIG1 PROTEIN PRECURSOR SIGNAL PD020194: M1-P38	BLAST_PRODOM

Table 3

SEQ ID NO:	Incyte Polypeptide ID	Amino Acid Residues	Signature Sequences, Domains and Motifs	Analytical Methods and Databases
			IMMUNOGLOBULIN DM00001|A53449|497-587: T392-S483 DM00001|A53449|405-495: A300-V391 DM00001|A53449|32-110: P32-S111 DM00001|A53449|126-206: T126-V207	BLAST_DOMO
			Potential Phosphorylation Sites: S67 S133 S164 S170 S184 S301 S331 S337 S344 S402 S407 S444 S453 S467 S483 S512 S573 S585 S608 S667 S692 S710 S712 T47 T244 T261 T365 T476 T543 T556 T649 T653 Y98	MOTIFS
			Potential Glycosylation Sites: N65 N193 N363 N384 N660	MOTIFS
17	7517783CD1	999	Signal Peptide: M2-G19, M2-Q24, M2-G18	HMMER
			Fibronectin type 3 domain: T774-G859, P598-G655, P874-G954, E672-G758	HMMER_SMART
			Immunoglobulin: K416-T497, P506-R595, P234-Y315, I324-V404, R129-R217, S35-A119	HMMER_SMART
			Immunoglobulin C-2 Type: A240-G304, A330-G393, Q422-G486, S41-G107, S512-D584, R135-T203	HMMER_SMART
			Fibronectin type III domain: P773-S862, P874-S957, P671-S761, P598-S658	HMMER_PFAM
			Immunoglobulin domain: G242-A299, G424-A481, G514-V579, G137-V198, D43-A102, E332-A388	HMMER_PFAM
			I type Ig domains from SCOP: A407-I502, K317-V404, T503-P601	HMMER_INCY
			Ig superfamily from SCOP: W320-D409, E230-N306, F28-A119, I502-P600, F410-T500, K125-P213	HMMER_INCY
			RECEPTOR INTERLEUKIN-1 PRECURSOR. PD02870: D510-D526, L371-V403, L464-E498, E235-A251, N603-K627	BLIMPS_PRODOM
			PRECURSOR SIGNAL CONTACTIN CELL ADHESION NEUROFASCIN GLYCOPROTEIN GP135 IMMUNOGLOBULIN FOLD PD001890: L659-P773	BLAST_PRODOM
			ADHESION PRECURSOR SIGNAL CELL IMMUNOGLOBULIN FOLD GLYCOPROTEIN GPIANCHOR REPEAT CONTACTIN PD005229: V865-I962	BLAST_PRODOM
			SIMILAR TO FIBRONECTIN TYPE III PD073047: N301-G560	BLAST_PRODOM
			NEURAL CELL ADHESION MOLECULE CLOSE HOMOLOGUE OF L1 L1LIKE PROTEIN PD066559: E482-G596	BLAST_PRODOM

Table 3

SEQ ID NO:	Incyte Polypeptide ID	Amino Acid Residues	Signature Sequences, Domains and Motifs	Analytical Methods and Databases
			IMMUNOGLOBULIN DM00001|A53449|497-587: T497-S588 DM00001|A53449|405-495: A405-V496 DM00001|A53449|32-110: P32-S111 DM00001|A53449|126-206: T126-V207	BLAST_DOMO
			Potential Phosphorylation Sites: S67 S133 S164 S170 S184 S270 S279 S342 S348 S377 S397 S406 S436 S442 S449 S507 S512 S549 S558 S572 S588 S649 S661 S684 S743 S768 S786 S788 S823 S834 T47 T244 T364 T470 T581 T619 T632 T725 T729 T868 T869 T926 T929 T955 T966 Y98	MOTIFS
			Potential Glycosylation Sites: N65 N193 N375 N468 N489 N736 N831 N866 N884 N902 N927	MOTIFS
18	7522607CD1	200	signal_cleavage: M1-S20	SPSCAN
			Signal Peptide: M1-T15, M1-S20, M1-G23, M1-L17	HMMER
			Furin-like repeats: M91-A135, A34-D85	HMMER_SMART
			TNFR/NGFR family cysteine-rich region IPB001368: C75-D85	BLIMPS_BLOCKS
			Disintegrin IPB001762: N92-P145	BLIMPS_BLOCKS
			Potential Phosphorylation Sites: S20 S33 S57 S169 S176 Y119	MOTIFS
			Potential Glycosylation Sites: N137	MOTIFS
19	7521142CD1	123	signal_cleavage: M1-A22	SPSCAN
			Signal Peptide: M5-A21, M5-A22, M1-A21, M1-C23, M5-C29, M1-C29, M1-A22	HMMER
			Leucine rich repeat N-terminal domain: A22-L55	HMMER_SMART
			Leucine-rich repeats, typical (most populated) subfamily: M74-G97, L98-Q121	HMMER_SMART
			Leucine Rich Repeat: N76-K99, K52-P75, Q100-Q121	HMMER_PFAM
			Leucine rich repeat N-terminal domain: A22-S50	HMMER_PFAM
			Leucine-rich repeat signature PR00019: L101-L114, L98-I111	BLIMPS_PRINTS
			CHONDROADHERIN PRECURSOR CARTILAGE LEUCINE-RICH PROTEIN BONE REPEAT SIGNAL PD151229: C23-Q100	BLAST_PRODOM
			LRR REPEAT DM00016|A53860|81-107: S79-L106	BLAST_DOMO
			Potential Phosphorylation Sites: S50 S70 S107	MOTIFS
20	7521689CD1	101	signal_cleavage: M1-A18	SPSCAN
			Signal Peptide: M1-A19, M1-F20, M1-V21	HMMER

Table 3

SEQ ID NO:	Incyte Polypeptide ID	Amino Acid Residues	Signature Sequences, Domains and Motifs	Analytical Methods and Databases
			MICROFIBRIL-ASSOCIATED GLYCOPROTEIN 3 PRECURSOR IMMUNOGLOBULIN FOLD EXTRACELLULAR MATRIX SIGNAL PD066579: M1-G99	BLAST_PRODOM
			IMMUNOGLOBULIN DM00001|P55082|55-134: H55-G99	BLAST_DOMO
			Potential Phosphorylation Sites: S46 S54 S56 S78	MOTIFS
			Potential Glycosylation Sites: N36 N41	MOTIFS
21	2878775CD1	1040	signal_cleavage: M1-G15	SPSCAN
			Signal Peptide: M1-G15, M1-S18	HMMER
			Cadherin repeats.: V485-P572, V603-Q686, E144-P248, V34-P120, L272-P356, F380-P461	HMMER_SMART
			Cadherin domain: Y255-L349, R587-V679, Y468-V565, V363-S454, L127-L241	HMMER_PFAM
			Cytosolic domain: R739-C1040 Transmembrane domain: L716-V738 Non-cytosolic domain: M1-T715	TMHMMER
			C-terminal cystine knot IPB000359: A753-G766, F727-A775	BLIMPS_BLOCKS
			Cadherins extracellular repeated domain signature: T432-I482	PROFILESCAN
			Cadherins extracellular repeated domain signature: V200-V269	PROFILESCAN
			Cadherins extracellular repeated domain signature: V323-V377	PROFILESCAN
			Cadherin signature PR00205: P62-R81, F250-E279, S428-E440, A442-P461, P461-E474, E521-D547, G556-A573	BLIMPS_PRINTS
			Connexin36 (Cx36) signature PR01131: G687-H699	BLIMPS_PRINTS
			GLYCOPROTEIN PROTEIN SPIKE E2 PRECURSOR PEPLOMER. PD00866: E279-A330, E424-N476, R546-D569, C760-T769	BLIMPS_PRODOM
			CELL ADHESION TRANSMEMBRANE CALCIUM BINDING REPEAT GLYCOPROTEIN KIAA0345LIKE PROTOCADHERIN PROTEIN PRECURSOR PD017893: Q19-E130	BLAST_PRODOM
			PROTOCADHERIN 3 PD176553: I574-P685	BLAST_PRODOM
			INSECTICIDAL TOXIN RECEPTOR BTR1 PRECURSOR RECEPTOR GLYCOPROTEIN TRANSMEMBRANE SIGNAL REPEAT CELL ADHESION PD131770: A222-I358 I449-Q686 K343-V565	BLAST_PRODOM
			PROTOCADHERIN 68 CELL ADHESION GLYCOPROTEIN TRANSMEMBRANE CALCIUMBINDING REPEAT PD131829: R704-S897	BLAST_PRODOM

Table 3

<table>
<tr><th>SEQ ID NO:</th><th>Incyte Polypeptide ID</th><th>Amino Acid Residues</th><th>Signature Sequences, Domains and Motifs</th><th>Analytical Methods and Databases</th></tr>
<tr><td></td><td></td><td></td><td>CADHERIN REPEAT
DM00030|P33450|1079-1181: E389-D491
DM00030|P33450|2196-2306: L421-E492
DM00030|P33450|2417-2519: N390-E492
DM00030|P34616|1682-1783: E389-D491</td><td>BLAST_DOMO</td></tr>
<tr><td></td><td></td><td></td><td>Potential Phosphorylation Sites: S49 S121 S145 S155 S258 S294 S304 S381 S408 S486 S501 S783 S829 S907 S921 S950 S980 S1011 T23 T173 T202 T237 T275 T307 T362 T383 T445 T488 T640 T710 T846 T926 T1035 Y598</td><td>MOTIFS</td></tr>
<tr><td></td><td></td><td></td><td>Potential Glycosylation Sites: N273 N557 N870 N876 N896 N939</td><td>MOTIFS</td></tr>
<tr><td></td><td></td><td></td><td>Cadherins extracellular repeated domain signature: I110-P120, I238-P248, V346-P356, V451-P461, I562-P572</td><td>MOTIFS</td></tr>
<tr><td></td><td></td><td></td><td>EGF-like domain signature 1: C759-C770</td><td>MOTIFS</td></tr>
<tr><td>22</td><td>7521207CD1</td><td>58</td><td>signal_cleavage: M1-A16</td><td>SPSCAN</td></tr>
<tr><td></td><td></td><td></td><td>Signal Peptide: M1-A16, M1-P18, M1-T21, M1-P22, M1-W25</td><td>HMMER</td></tr>
<tr><td></td><td></td><td></td><td>Cytosolic domain: K24-D58
Transmembrane domain: W4-L23
Non-cytosolic domain: M1-T3</td><td>TMHMMER</td></tr>
<tr><td></td><td></td><td></td><td>Osteonectin domain IPB001999: A14-D58</td><td>BLIMPS_BLOCKS</td></tr>
<tr><td></td><td></td><td></td><td>Osteopontin IPB002038: M1-R30</td><td>BLIMPS_BLOCKS</td></tr>
<tr><td></td><td></td><td></td><td>AMELOGENIN ISOFORM PRECURSOR EXTRACELLULAR MATRIX PHOSPHORYLATION ENAMEL REPEAT SIGNAL ALTERNATIVE PD007277: M1-V19 V19-P35</td><td>BLAST_PRODOM</td></tr>
<tr><td></td><td></td><td></td><td>AMELOGENIN DM05282|S50218|10-74: V19-D58</td><td>BLAST_DOMO</td></tr>
<tr><td></td><td></td><td></td><td>Potential Phosphorylation Sites: S28 T49 T52</td><td>MOTIFS</td></tr>
<tr><td>23</td><td>7521283CD1</td><td>74</td><td>signal_cleavage: M1-A16</td><td>SPSCAN</td></tr>
<tr><td></td><td></td><td></td><td>Signal Peptide: M1-A16, M1-P18, M1-P21, M1-P23, M1-G24, M1-P26</td><td>HMMER</td></tr>
<tr><td></td><td></td><td></td><td>Osteonectin domain IPB001999: T3-P18</td><td>BLIMPS_BLOCKS</td></tr>
<tr><td></td><td></td><td></td><td>AMELOGENIN ISOFORM PRECURSOR EXTRACELLULAR MATRIX PHOSPHORYLATION ENAMEL REPEAT SIGNAL ALTERNATIVE PD007277: M1-P51</td><td>BLAST_PRODOM</td></tr>
<tr><td></td><td></td><td></td><td>AMELOGENIN DM05282|S50218|10-74: L10-D74</td><td>BLAST_DOMO</td></tr>
</table>

Table 3

SEQ ID NO:	Incyte Polypeptide ID	Amino Acid Residues	Signature Sequences, Domains and Motifs	Analytical Methods and Databases
			Potential Phosphorylation Sites: S44 T65 T68	MOTIFS
			Potential Glycosylation Sites: N30	MOTIFS
24	7522210CD1	366	signal_cleavage: M1-A26	SPSCAN
			Signal Peptide: G9-A26, M1-N20, P8-P30, M1-A26, M1-P30, M1-A26, P4-A26	HMMER
			C-type lectin (CTL) or carbohydrate-recognition domain: F217-Q353	HMMER_SMART
			Lectin C-type domain: E236-F354	HMMER_PFAM
			EGF-like domain IPB000561: C169-G177	BLIMPS_BLOCKS
			Extracellular proteins SCP/Tpx-1/Ag5/PR-1/Sc7 IPB001283: D68-G84, A97-N107	BLIMPS_BLOCKS
			C-type lectin domain signature and profile: E308-S366	PROFILESCAN
			Type II EGF-like signature PR00010: G208-V215, S165-Y175, A141-V147	BLIMPS_PRINTS
			Laminin G domain proteins. PF00054: P173-C184	BLIMPS_PFAM
			Sushi domain proteins (SCR repeat proteins. PF00084: C169-P173, I164-Y175, G106-I115	BLIMPS_PFAM
			C-TYPE LECTIN DM00035|P22897|639-777: E236-Q353	BLAST_DOMO
			Potential Phosphorylation Sites: S48 S235 S259 S282 S301 T24 T93 T176 T275 T280 T294 T297	MOTIFS
			Potential Glycosylation Sites: N163	MOTIFS
			C-type lectin domain signature: C328-C352	MOTIFS
			EGF-like domain signature 1: C169-C180, C200-C211	MOTIFS
			EGF-like domain signature 2: C169-C180, C200-C211	MOTIFS
25	7519488CD1	74	signal_cleavage: M1-S22	SPSCAN
			Signal Peptide: M1-P24	HMMER
			Cytosolic domain: P38-T74 Transmembrane domain: L15-L37 Non-cytosolic domain: M1-Q14	TMHMMER
			ACETYLCHOLINESTERASE ASSOCIATED COLLAGEN PD060058: M1-P73	BLAST_PRODOM
			ACETYLCHOLINESTERASE COLLAGENIC TAIL PEPTIDE PRECURSOR ACHE Q SUBUNIT SIGNAL SYNAPSE NEUROTRANSMITTER DEGRADATION REPEAT PD091309: P6-F67	BLAST_PRODOM
26	7519965CD1	272	signal_cleavage: M1-S22	SPSCAN
			Signal Peptide: M1-P24	HMMER
			Collagen triple helix repeat (20 copies): G150-K209, S91-K149, G210-P269	HMMER_PFAM

Table 3

SEQ ID NO:	Incyte Polypeptide ID	Amino Acid Residues	Signature Sequences, Domains and Motifs	Analytical Methods and Databases
			ACETYLCHOLINESTERASE COLLAGENIC TAIL PEPTIDE PRECURSOR ACHE Q SUBUNIT SIGNAL SYNAPSE NEUROTRANSMITTER DEGRADATION REPEAT PD091309: P6-P128	BLAST_PRODOM
			ACETYLCHOLINESTERASE ASSOCIATED COLLAGEN PD060058: M1-L98	BLAST_PRODOM
			COLLAGEN ALPHA PRECURSOR CHAIN REPEAT SIGNAL CONNECTIVE TISSUE EXTRACELLULAR MATRIX PD000007: G99-D184, P170-P260	BLAST_PRODOM
			PRECOLLAGEN P PRECURSOR SIGNAL PD072959: G162-P257, S89-G180, G186-G270	BLAST_PRODOM
			FIBRILLAR COLLAGEN CARBOXYL-TERMINAL DM00019|Q03637|147-282: G120-P251, M94-G216 DM00019|P02461|924-1148: G96-P259 DM00019|P04258|434-662: P57-A262 DM00019|A41182|873-1104: S89-G258	BLAST_DOMO
			Potential Phosphorylation Sites: S172 S178 T221	MOTIFS
27	7519985CD1	82	signal_cleavage: M1-A18	SPSCAN
			Signal Peptide: M1-G16, M1-A18, M1-N20, M1-P23, M1-L21	HMMER
			UROPLAKIN III UROPLKIN PRECURSOR UPIII GLYCOPROTEIN TRANSMEMBRANE SIGNAL PD127194: M1-I77	BLAST_PRODOM
28	7520002CD1	77	signal_cleavage: M1-A16	SPSCAN
			Signal Peptide: M1-A16, M1-P18, M1-Q21, M1-G23	HMMER
			DECORIN BONE PROTEOGLYCAN II PRECURSOR PGS2 GLYCOPROTEIN CONNECTIVE TISSUE EXTRACELLULAR PD009349: M1-D45	BLAST_PRODOM
29	7520014CD1	195	signal_cleavage: M1-Q15	SPSCAN
			Signal Peptide: M1-Q15, M1-G18, M1-A20, M1-P23, M1-S21, M1-T19	HMMER
			OSTEOINDUCTIVE; LB; PROTEOGLYCAN; DM08598|P19879|61-128: N122-F162 DM08598|A41781|62-141: S88-Y157	BLAST_DOMO
			Potential Phosphorylation Sites: S63 S188	MOTIFS
30	7520039CD1	168	Signal Peptide: M1-E24	HMMER

Table 3

SEQ ID NO:	Incyte Polypeptide ID	Amino Acid Residues	Signature Sequences, Domains and Motifs	Analytical Methods and Databases
			Cytosolic domain: F141-P168 Transmembrane domain: V118-I140 Non-cytosolic domain: M1-M117	TMHMMER
			Intercellular adhesion molecule/vascular cell adhesion molecule-1 signature PR01472: K35-T51	BLIMPS_PRINTS
			PRECURSOR SIGNAL ADHESION TRANSMEMBRANE INTERCELLULAR IMMUNOGLOBULIN FOLD CELL GLYCOPROTEIN REPEAT PD005863: P20-P111 V106-C138	BLAST_PRODOM
			INTERCELLULAR ADHESION ICAM2 MOLECULE 2 PRECURSOR CD102 IMMUNOGLOBULIN FOLD CELL GLYCOPROTEIN PD024043: F139-P168	BLAST_PRODOM
			INTERCELLULAR ADHESION MOLECULE DM01682|P13598|1-113: M1-P111 DM01682|P35330|1-113: M1-P111 DM01682|P32942|8-118: L16-Y109	BLAST_DOMO
			Potential Phosphorylation Sites: S96 T62	MOTIFS
			Potential Glycosylation Sites: N47 N82 N105	MOTIFS
31	7520053CD1	87	signal_cleavage: M1-A18	SPSCAN
			Signal Peptide: M1-G16, M1-A18, M1-N20, M1-P23, M1-L21	HMMER
			UROPLAKIN III UROPLKIN PRECURSOR UPIII GLYCOPROTEIN TRANSMEMBRANE SIGNAL PD127194: M1-S69	BLAST_PRODOM
32	7523262CD1	207	signal_cleavage: M1-A16	SPSCAN
			Signal Peptide: M1-A16, M1-P18, M1-G21, M1-A23, M1-G25, M1-S22	HMMER
			Cytosolic domain: R27-R207 Transmembrane domain: L4-R26 Non-cytosolic domain: M1-S3	TMHMMER
			Intercellular adhesion molecule/vascular cell adhesion molecule-1 signature PR01472: F56-S72	BLIMPS_PRINTS
			PRECURSOR SIGNAL ADHESION TRANSMEMBRANE INTERCELLULAR IMMUNOGLOBULIN FOLD CELL GLYCOPROTEIN REPEAT PD005863: L39-T204	BLAST_PRODOM
			LANDSTEINERWIENER BLOOD GROUP GLYCOPROTEIN PRECURSOR LW PROTEIN IMMUNOGLOBULIN FOLD CELL ADHESION TRANSMEMBRANE SIGNAL ALTERNATIVE SPLICING POLYMORPHISM ANTIGEN PD124387: M1-A40	BLAST_PRODOM

Table 3

SEQ ID NO:	Incyte Polypeptide ID	Amino Acid Residues	Signature Sequences, Domains and Motifs	Analytical Methods and Databases
			INTERCELLULAR ADHESION MOLECULE DM01682|I59300|17-132: R30-P134 DM01682|I48950|4-120: P47-P134	BLAST_DOMO
			Potential Phosphorylation Sites: S33 S80 S190 T28 T90 T125 T175	MOTIFS
			Potential Glycosylation Sites: N68 N78	MOTIFS
33	7523270CD1	259	signal_cleavage: M1-G17	SPSCAN
			Signal Peptide: M1-G17, M1-P19, M1-G23	HMMER
			Immunoglobulin: E55-R164	HMMER_SMART
			Immunoglobulin V-Type: S65-V148	HMMER_SMART
			Link (Hyaluronan-binding): R164-R257	HMMER_SMART
			Extracellular link domain: G165-Y256	HMMER_PFAM
			Immunoglobulin domain: G63-V148	HMMER_PFAM
			V type Ig domains from SCOP (antibody variant): L49-L176	HMMER_INCY
			Link domain IPB000538: A181-P233	BLIMPS_BLOCKS
			Vascular cell adhesion molecule-1 (VCAM-1) signature PR01474: R137-D150	BLIMPS_PRINTS
			GLYCOPROTEIN PRECURSOR PROTEIN PROTEOGLYCAN SIGNAL REPEAT CORE EGF-LIKE DOMAIN IMMUNOGLOBULIN PD000918: K175-Y256	BLAST_PRODOM
			COMPLEMENT FACTOR H REPEAT DM00260|A54423|254-355: E162-R257 DM00260|P55066|256-358: E162-H259 DM00260|P55252|256-353: G165-H259 DM00260|S39796|617-718: L163-R257	BLAST_DOMO
			Potential Phosphorylation Sites: S29 S80 S111 S219 T53 T177 Y144	MOTIFS
			Link: C185-C231	MOTIFS
34	7523287CD1	168	signal_cleavage: M1-A29	SPSCAN
			Signal Peptide: P5-A29, M1-A29, P7-A29, T10-A29	HMMER
			Immunoglobulin: A37-L150	HMMER_SMART
			Potential Phosphorylation Sites: S110 S141 T93	MOTIFS
			Potential Glycosylation Sites: N108	MOTIFS

Table 3

SEQ ID NO:	Incyte Polypeptide ID	Amino Acid Residues	Signature Sequences, Domains and Motifs	Analytical Methods and Databases
35	7521825CD1	373	signal_cleavage: M1-S23	SPSCAN
			Signal Peptide: M1-A19, M1-G20, M1-A22, M1-S23, M1-G28, M1-S30, M1-Y24	HMMER
			Complement component C1q domain: P221-D367	HMMER_SMART
			C1q domain: A229-L364	HMMER_PFAM
			Collagen triple helix repeat (20 copies): G168-Q227	HMMER_PFAM
			Complement C1q protein IPB001073: G174-G198, T243-V289, V327-D343, T357-G366	BLIMPS_BLOCKS
			Complement C1Q domain signature PR00007: P237-V263, F264-H283, T324-V345, P355-Y365	BLIMPS_PRINTS
			COLLAGEN ALPHA PRECURSOR CHAIN REPEAT SIGNAL CONNECTIVE TISSUE EXTRACELLULAR MATRIX PD000007: G138-P224, G168-P226, P133-E214, S132-G216	BLAST_PRODOM
			PRECOLLAGEN P PRECURSOR SIGNAL PD072959: S141-P224	BLAST_PRODOM
			SIMILAR TO CUTICULAR COLLAGEN PD067228: C129-G216, E131-Q227	BLAST_PRODOM
			FIBRILLAR COLLAGEN CARBOXYL-TERMINAL DM00019|P18834|155-317: G34-S54, G123-P221, L137-G210, P133-P226, P146-P224, R119-P226 DM00019|P20908|1300-1524: G123-P226, P133-P267, Q122-P221, Q122-Q227, R27-Q45, Q122-P226 DM00019|P34391|91-258: C120-A223, G125-Q227, G138-P224, G309-P322, T16-Q47, G123-G219	BLAST_DOMO
			FIBRILLAR COLLAGEN CARBOXYL-TERMINAL DM00019|S18803|1304-1528: G123-P226, L167-P224, P133-P267, Q122-P221, Q122-Q227, R27-Q45, Q122-P226	BLAST_DOMO
			Potential Phosphorylation Sites: T112 T114 T143	MOTIFS
			Potential Glycosylation Sites: N154	MOTIFS
			Leucine_Zipper: L349-L370	MOTIFS
36	7521844CD1	237	signal_cleavage: M1-G17	SPSCAN
			Signal Peptide: M1-G19	HMMER
			Domain first found in C1r, C1s, uEGF, and bone morphogenetic protein: C135-M207	HMMER_SMART
			Link (Hyaluronan-binding): A34-N129	HMMER_SMART
			CUB domain: C135-Y204	HMMER_PFAM
			Extracellular link domain: A35-Y128	HMMER_PFAM
			Link domain IPB000538: E53-F105	BLIMPS_BLOCKS
			Link module signature PR01265: K46-C58, A74-M87, V92-C103, W123-A132	BLIMPS_PRINTS

Table 3

SEQ ID NO:	Incyte Polypeptide ID	Amino Acid Residues	Signature Sequences, Domains and Motifs	Analytical Methods and Databases
			LAMININ CHAIN EGF-LIKE DOMAIN PD00320: G36-L49	BLIMPS_PRODOM
			GLYCOPROTEIN PRECURSOR PROTEIN PROTEOGLYCAN SIGNAL REPEAT CORE EGFLIKE DOMAIN IMMUNOGLOBULIN PD000918: V37-Y128	BLAST_PRODOM
			PROTEIN TUMOR NECROSIS FACTOR INDUCIBLE TSG6 PRECURSOR HYALURONATE BINDING CELL ADHESION PD151779: M1-G36	BLAST_PRODOM
			PROTEIN TUMOR NECROSIS FACTOR INDUCIBLE TSG6 PRECURSOR HYALURONATE BINDING CELL ADHESION PD019802: V205-L237	BLAST_PRODOM
			COMPLEMENT FACTOR H REPEAT DM00260|P98066|31-129: L31-P130 DM00260|P07897|482-582: V37-Y128 DM00260|P13611|145-245: V37-Y128	BLAST_DOMO
			C1R/C1S REPEAT DM00162|P98066|131-247: H131-Y169 K168-M207	BLAST_DOMO
			Potential Phosphorylation Sites: S44 S120 T50 T67 T186	MOTIFS
			Potential Glycosylation Sites: N118 N218	MOTIFS
			Link: C58-C103	MOTIFS
37	7521864CD1	163	signal_cleavage: M1-A16	SPSCAN
			Signal Peptide: M1-A16, M1-S18, M1-S20	HMMER
			COLLAGEN ALPHA PRECURSOR CHAIN REPEAT SIGNAL CONNECTIVE TISSUE EXTRACELLULAR MATRIX PD000007: G34-P132, G65-P161, P50-Q157	BLAST_PRODOM
			INTERMEDIATE CHAIN MULTIPLE BANDED ANTIGEN PRECURSOR SIGNAL PD103341: G34-Q157, G38-P161, R40-Q163	BLAST_PRODOM
			PROTEIN REPEAT MICROTUBULE-ASSOCIATED MICROTUBULES PHOSPHORYLATION BASSOON ALTERNATIVE SPLICING LARGE PROLINERICH PD005493: P36-P161, P36-Q159, P41-P162, Q46-P162, S33-Q159	BLAST_PRODOM
			PROLINE-RICH PROTEIN DM01281|P04280|17-124: E17-G126, N45-P145, N66-P162 DM01281|P04280|212-315: G34-G142, P50-G149 DM01281|P10163|107-216: G34-Q133, G44-P155, G65-P162	BLAST_DOMO

Table 3

SEQ ID NO:	Incyte Polypeptide ID	Amino Acid Residues	Signature Sequences, Domains and Motifs	Analytical Methods and Databases
			PROLINE-RICH PROTEIN DM01281|P10163|17-105: E17-Q105, G34-G64, G55-G126, G76-P145, G97-P162	BLAST_DOMO
			Potential Phosphorylation Sites: S14 S18 S24 S33 S91	MOTIFS
			Potential Glycosylation Sites: N66 N87	MOTIFS
38	7522020CD1	207	C-type lectin (CTL) or carbohydrate-recognition domain: C117-A207	HMMER_SMART
			Lectin C-type domain: E134-F206	HMMER_PFAM
			Cytosolic domain: E98-A207 Transmembrane domain: V75-L97 Non-cytosolic domain: M1-E74	TMHMMER
			C-type lectin domain IPB001304: W121-C145, W172-W184	BLIMPS_BLOCKS
			PROTEIN RECEPTOR TYPE II INTEGRAL MEMBRANE TRANSMEMBRANE MULTIGENE FAMILY SIGNAL ANCHOR PD009256: M1-G77	BLAST_PRODOM
			PROTEIN RECEPTOR TYPE II INTEGRAL MEMBRANE TRANSMEMBRANE MULTIGENE FAMILY SIGNAL ANCHOR PD000827: T146-H194	BLAST_PRODOM
			C-TYPE LECTIN DM00035|P26717|110-228: K110-H194 DM00035|P26715|112-230: K110-H194 DM00035|I54524|110-229: K110-H194 DM00035|Q07108|78-195: H113-W197	BLAST_DOMO
			Potential Phosphorylation Sites: S32 S201 T7 T34 T108 T137 T146 T196	MOTIFS
			Potential Glycosylation Sites: N27 N100 N149 N178	MOTIFS
39	758410CD1	531	signal_cleavage: M1-A50	SPSCAN
			Signal Peptide: T30-C51, M1-A50	HMMER
			Collagen triple helix repeat (20 copies): G312-S371, G175-K234, G457-C516, G235-Q294, L398-P456, C116-A174	HMMER_PFAM
			COLLAGEN ALPHA PRECURSOR CHAIN REPEAT SIGNAL CONNECTIVE TISSUE EXTRACELLULAR MATRIX PD000007: G175-G277, G312-E385, G312-G412, P120-E215, P129-G229, P156-G244, P203-D296, P338-E437, P402-G496, P405-E506, P408-G511	BLAST_PRODOM
			SIMILAR TO CUTICULAR COLLAGEN PD067228: A376-G481, G312-G412, P179-E272, P243-E331, P313-G415, P316-D426, P407-G499, P446-P526, Q201-E293, R110-G229, R131-D237	BLAST_PRODOM

Table 3

SEQ ID NO:	Incyte Polypeptide ID	Amino Acid Residues	Signature Sequences, Domains and Motifs	Analytical Methods and Databases
			COLLAGEN ALPHA 1(IV) CHAIN DM04676|A54122|1-681: G79-G250, L100-G527, P119-P524, P120-G505, P120-K480, Q170-G527, R110-P526, R126-G527 DM04676|P08572|10-585: G124-D513, G130-P526, G193-V525, P119-F473, P119-G502, P120-G511, R110-G522 DM04676|P53420|18-722: E111-L523, G175-G527, P119-G522, P119-G527, P120-G511, P120-P526, S114-P524, V117-G448	BLAST_DOMO
			FIBRILLAR COLLAGEN CARBOXYL-TERMINAL DM00019|P12107|894-1149: E111-G354, G160-G439 G268-G511, G283-G522, G312-P526, G321-P524, P119-S386, R107-G289	BLAST_DOMO
			Potential Phosphorylation Sites: S141 S386 S390 S484 T108 T192 T458	MOTIFS
			Leucine_Zipper: L54-L75	MOTIFS
			Rgd: R132-D134	MOTIFS
40	7520759CD1	347	signal_cleavage: M1-S27	SPSCAN
			Signal Peptide: P9-S27, Q7-S27, Q7-P29, M1-G26, M1-S27, G5-S27	HMMER
			Thrombospondin type 1 repeats: W139-P191, W80-P134	HMMER_SMART
			Thrombospondin type 1 domain: S140-C190, S81-C133	HMMER_PFAM
			PRECURSOR GLYCOPROTEIN SIGNAL RECEPTOR PD01719: W139-P166	BLIMPS_PRODOM
			PROPERDIN PRECURSOR SIGNAL COMPLEMENT ALTERNATE PATHWAY GLYCOPROTEIN REPEAT DISEASE MUTATION PD012461: P294-E346	BLAST_PRODOM
			PROPERDIN PRECURSOR SIGNAL COMPLEMENT ALTERNATE PATHWAY GLYCOPROTEIN REPEAT DISEASE MUTATION PD012025: A24-W80	BLAST_PRODOM
			PRECURSOR SIGNAL PROTEIN PROPERDIN SEMAPHORIN F G COMPLEMENT ALTERNATE PATHWAY PD017722: V189-P293, D129-228, C72-170	BLAST_PRODOM
			PROTEIN SIGNAL PRECURSOR REPEAT GLYCOPROTEIN THROMBOSPONDIN CIRCUMSPOROZOITE CELL ADHESION MALARIA PD000485: W139-C190	BLAST_PRODOM

Table 3

SEQ ID NO:	Incyte Polypeptide ID	Amino Acid Residues	Signature Sequences, Domains and Motifs	Analytical Methods and Databases
			THROMBOSPONDIN TYPE 1 REPEAT DM00275|P27918|119-176: T119-Q177 DM00275|P27918|365-419: Q243-M298 DM00275|P27918|300-363: A178-Q242 DM00275|P27918|65-117: Q65-G118	BLAST_DOMO
			Potential Phosphorylation Sites: S40 S52 S77 S112 S198 S214 S259 S297 T25 T34 T92 T158 T227	MOTIFS
			Potential Glycosylation Sites: N306	MOTIFS
			Growth factor and cytokines receptors family signature 2: G258-S264	MOTIFS
41	7522915CD1	366	signal_cleavage: M1-T34	SPSCAN
			Signal Peptide: M1-T34, M1-R36	HMMER
			Epidermal growth factor-like domain.: E137-T175, V92-L133	HMMER_SMART
			Calcium-binding EGF-like domain: D134-T175, D89-L133	HMMER_SMART
			EGF-like domain: C138-C174, C93-K126	HMMER_PFAM
			Calcium-binding EGF-like domain IPB001881: C111-R121, C150-C161	BLIMPS_BLOCKS
			Complement Clq protein IPB001073: G247-G271	BLIMPS_BLOCKS
			Type I EGF signature PR00009: L133-H148, L154-Y165	BLIMPS_PRINTS
			COLLAGEN ALPHA PRECURSOR CHAIN REPEAT SIGNAL CONNECTIVE TISSUE EXTRACELLULAR MATRIX PD000007: P246-E341, G254-P355	BLAST_PRODOM
			PRECOLLAGEN P PRECURSOR SIGNAL PD072959: N242-P329, G256-P351	BLAST_PRODOM
			PRECURSOR SIGNAL COLLAGEN ALPHA 3IX CHAIN EXTRACELLULAR MATRIX CONNECTIVE TISSUE PD028299: G247-R326, G250-G327	BLAST_PRODOM
			SIMILAR TO CUTICULAR COLLAGEN PD067228: P246-G327, P248-G327, G235-G315, P257-P355	BLAST_PRODOM
			FIBRILLAR COLLAGEN CARBOXYL-TERMINAL DM00019|P29400|705-886: P246-P361, L245-P361, P246-P355, T243-G340 DM00019|P18834|155-317: P223-P336, G247-P351, P252-P355, G247-P355, G265-T354 DM00019|Q02388|1373-1547: P3-G17, L245-E341, G247-G327, T243-G327, G250-R342, P246-P311 DM00019|P39061|455-643: S241-W356, P246-P355, G247-G340, D228-P325	BLAST_DOMO

Table 3

SEQ ID NO:	Incyte Polypeptide ID	Amino Acid Residues	Signature Sequences, Domains and Motifs	Analytical Methods and Databases
			Potential Phosphorylation Sites: S62 S156 S316 S338 S347 T34 T54 T175 T184 T201 Y131	MOTIFS
			Potential Glycosylation Sites: N142 N182	MOTIFS
			Cell attachment sequence: R176-D178	MOTIFS
			EGF-like domain signature 2: C159-C174	MOTIFS
			Calcium-binding EGF-like domain pattern signature: D134-C159	MOTIFS
42	7522936CD1	247	signal_cleavage: M1-S22	SPSCAN
			Signal Peptide: M1-P24	HMMER
			Collagen triple helix repeat (20 copies): G98-K157, G176-P235	HMMER_PFAM
			ACETYLCHOLINESTERASE-ASSOCIATED COLLAGEN PD060058: M1-P73	BLAST_PRODOM
			COLLAGEN ALPHA PRECURSOR CHAIN REPEAT SIGNAL CONNECTIVE TISSUE EXTRACELLULAR MATRIX PD000007: G69-G158, P56-E150, P145-P235, G98-G191, P133-G221	BLAST_PRODOM
			PRECURSOR SIGNAL COLLAGEN ALPHA 3IX CHAIN EXTRACELLULAR MATRIX CONNECTIVE TISSUE PD028299: G80-G164, R148-P234, K124-G209, G140-G215, G95-G185, G113-G197	BLAST_PRODOM
			ACETYLCHOLINESTERASE COLLAGENIC TAIL PEPTIDE PRECURSOR ACHE Q SUBUNIT SIGNAL SYNAPSE NEUROTRANSMITTER DEGRADATION REPEAT PD091309: P6-G140	BLAST_PRODOM
			FIBRILLAR COLLAGEN CARBOXYL-TERMINAL DM00019|Q03637|147-282: P94-P226, P65-G191 DM00019|Q07092|1216-1417: P58-P234, P65-P234, P73-P235, P60-G224, G74-G246, G98-G233, G116-P235, G146-A237 DM00019|P04258|434-662: T55-A237, P57-G245, P65-D238, G69-G245, G86-A247, G74-G221, P64-G203 DM00019|P02461|924-1148: P73-P234, P57-G233, P59-P234, G89-P234, G140-G246	BLAST_DOMO
			Potential Phosphorylation Sites: S147 S153 T196	MOTIFS

Table 4

Polynucleotide SEQ ID NO:/ Incyte ID/ Sequence Length	Sequence Fragments
43/7513225CB1 4720	1-281, 2-296, 2-867, 2-877, 3-261, 3-481, 3-630, 3-832, 3-842, 3-853, 3-4716, 4-239, 4-658, 4-702, 7-279, 8-593, 20-268, 33-560, 37-145, 86-494, 284-818, 365-743, 402-671, 422-939, 579-1139, 701-1582, 828-1512, 1130-2020, 1210-1412, 1306-1836, 1431-1990, 1575-2019, 1636-2261, 1668-2093, 1668-2258, 1676-1936, 1846-2108, 1862-2387, 1902-2554, 1936-2259, 1937-2459, 2028-2276, 2169-2490, 2177-2278, 2430-2720, 2430-2861, 2510-2861, 2586-3153, 2586-3172, 2598-3008, 2629-2912, 2629-3250, 2635-3188, 2712-3152, 2764-2993, 2800-3266, 2838-3302, 2867-3287, 2916-3097, 2954-3183, 2985-3281, 3017-3531, 3049-3368, 3076-3338, 3084-3572, 3084-3768, 3188-3452, 3210-3434, 3210-3836, 3253-3825, 3253-3847, 3253-3917, 3266-3629, 3290-3996, 3332-3476, 3351-3928, 3362-3641, 3382-3669, 3394-3936, 3403-3606, 3428-3936, 3431-4000, 3440-3865, 3501-3619, 3520-3945, 3534-3753, 3595-3850, 3605-3811, 3611-3921, 3625-4044, 3633-4127, 3633-4286, 3640-3903, 3646-3928, 3650-3882, 3652-3894, 3653-3910, 3658-3900, 3659-3975, 3661-4242,
	3669-4075, 3673-4165, 3674-4172, 3687-3827, 3687-3922, 3703-4308, 3715-3944, 3718-4204, 3730-3974, 3737-4256, 3738-4171, 3739-4023, 3747-3983, 3759-3914, 3763-3906, 3763-4340, 3771-4313, 3774-4395, 3793-4412, 3796-4071, 3803-3971, 3809-4423, 3829-4092, 3829-4460, 3831-4112, 3833-4105, 3849-4382, 3865-4262, 3868-4104, 3872-4476, 3878-4125, 3883-4097, 3885-4315, 3885-4385, 3892-4178, 3895-4160, 3902-4209, 3903-4171, 3908-4352, 3911-4254, 3917-4136, 3918-4205, 3921-4216, 3923-4173, 3927-4553, 3948-4232, 3952-4243, 3969-4236, 3973-4595, 4003-4233, 4037-4315, 4039-4612, 4043-4716, 4054-4321, 4057-4271, 4070-4704, 4077-4704, 4081-4362, 4083-4646, 4088-4703, 4089-4336, 4093-4341, 4097-4373, 4099-4706, 4100-4681, 4110-4688, 4114-4644, 4119-4706, 4126-4694, 4128-4609, 4132-4706, 4137-4716, 4138-4484, 4140-4668, 4140-4704, 4144-4393, 4158-4432, 4158-4705, 4167-4608, 4171-4720, 4182-4675, 4183-4425, 4187-4496, 4205-4559, 4210-4485, 4223-4453, 4223-4499, 4233-4611, 4250-4456, 4251-4521, 4252-4720, 4254-4720, 4260-4614,
	4260-4716, 4262-4514, 4263-4569, 4264-4720, 4268-4716, 4272-4528, 4273-4546, 4274-4720, 4278-4710, 4281-4524, 4284-4716, 4285-4720, 4294-4716, 4295-4720, 4304-4720, 4306-4704, 4318-4716, 4323-4571, 4323-4583, 4323-4716, 4327-4591, 4332-4631, 4345-4720, 4350-4716, 4358-4720, 4369-4593, 4376-4611, 4378-4720, 4381-4720, 4391-4572, 4391-4668, 4420-4682, 4420-4720, 4427-4663, 4445-4661, 4445-4693, 4445-4720, 4453-4677, 4464-4716, 4467-4716, 4470-4720, 4485-4654, 4485-4701, 4495-4720, 4515-4625, 4515-4720, 4562-4720, 4575-4716, 4607-4720

Table 4

Polynucleotide SEQ ID NO:/ Incyte ID/ Sequence Length	Sequence Fragments
44/7513288CB1 6633	1-6633, 122-573, 458-920, 525-782, 533-1125, 562-1110, 613-1086, 616-813, 621-1205, 734-785, 740-804, 749-1004, 749-1009, 825-1060, 855-1093, 1022-1085, 1022-1086, 1086-1346, 1126-1345, 1185-1292, 1185-1737, 1208-1562, 1281-1461, 1281-1499, 1282-1485, 1308-1391, 1317-1527, 1371-1478, 1371-1923, 1388-1436, 1388-1439, 1394-1457, 1394-1709, 1537-1592, 1654-1857, 1663-1737, 1723-1778, 1731-2393, 1816-1871, 2807-2829, 2826-3254, 3216-3466, 3344-3577, 3344-3579, 3369-3631, 3371-3616, 3379-3494, 3379-3764, 3381-3684, 3410-3929, 3468-3770, 3542-3613, 3542-3616, 3652-3734, 3652-3767, 3652-4037, 3654-3957, 3699-4037, 3868-4036, 3917-4012, 3917-4113, 3917-4162, 3927-4162, 3956-4162, 4015-4162, 4088-4548, 4088-4868, 4143-4433, 4180-4427, 4595-4915, 4601-4836, 4601-4858, 5728-6056, 5884-6383, 6157-6594
45/7513607CB1 1476	1-770, 1-806, 691-1476, 740-1476
46/7513991CB1 839	1-839, 135-677, 346-677, 348-677
47/7513298CB1 1488	1-175, 15-1334, 30-175, 59-275, 175-293, 175-399, 175-409, 175-418, 175-442, 175-458, 175-505, 175-622, 175-752, 176-425, 180-465, 183-355, 183-472, 185-401, 186-451, 189-669, 197-467, 197-468, 197-470, 199-727, 200-603, 203-361, 212-1055, 223-804, 230-484, 230-489, 232-961, 237-780, 241-869, 246-503, 248-452, 249-642, 250-505, 251-541, 251-550, 252-516, 254-908, 255-544, 257-451, 257-498, 258-845, 258-899, 259-508, 263-386, 263-475, 264-957, 266-472, 266-691, 270-512, 270-537, 271-827, 275-513, 275-954, 276-580, 280-468, 280-778, 281-454, 281-940, 282-537, 290-873, 292-687, 293-765, 297-571, 299-569, 313-881, 314-892, 318-848, 324-579, 324-898, 327-555, 335-591, 337-876, 340-948, 341-765, 346-603, 348-1027, 352-1120, 356-647, 374-694, 376-1085, 378-1000, 379-615, 380-999, 383-724, 390-996, 391-988, 392-1068, 397-672, 397-676, 397-992, 399-671, 408-656, 408-724, 410-971, 412-803, 412-923, 419-944, 425-666, 425-677, 428-694, 431-699, 431-706, 437-682, 438-1253, 439-766, 442-945, 443-989, 455-700, 455-1093,
	457-759, 458-1033, 460-765, 469-748, 470-741, 473-999, 475-731, 475-1267, 480-1072, 483-687, 483-1042, 484-1096, 488-723, 489-751, 491-719, 491-746, 494-1111, 494-1185, 496-723, 497-1129, 498-734, 499-730, 499-778, 500-1081, 500-1125, 500-1154, 502-776, 504-1169, 506-685, 506-1085, 508-1021, 509-808, 512-613, 512-670, 512-769, 512-784, 512-793, 516-794, 519-794, 524-772, 524-782, 525-685, 527-793, 530-1135, 531-1209, 533-1157, 536-770, 536-774, 536-781, 536-795, 541-827, 543-1132, 545-851, 545-1096, 552-817, 552-1006, 552-1086, 553-1079, 554-1085, 563-800, 563-859, 567-817, 567-833, 572-802, 573-825, 574-1203, 577-1251, 579-1292, 582-840, 582-869, 583-1176, 583-1199, 584-845, 584-860, 587-850, 588-897, 590-818, 590-838, 590-839, 595-894, 596-782, 598-885, 598-918, 598-921, 600-825, 600-879, 600-889, 607-1256, 608-993, 616-848, 616-1226, 622-1249, 626-1259, 626-1277, 629-1196, 643-1196, 643-1239, 647-856, 650-1151, 650-1240, 651-1265, 652-1274, 658-780, 658-960, 659-906, 661-1166, 661-1259, 662-891, 662-1300,

Table 4

Polynucleotide SEQ ID NO:/ Incyte ID/ Sequence Length	Sequence Fragments
	664-1195, 666-1086, 669-936, 669-1173, 670-1088, 671-943, 672-1222, 680-770, 681-1206, 683-1099, 686-1295, 688-960, 695-1258, 701-858, 701-950, 701-972, 703-1039, 709-1109, 710-894, 710-920, 710-923, 711-1271, 721-1166, 722-1059, 725-1308, 731-969, 731-1009, 731-1011, 731-1227, 734-1321, 738-1209, 740-1036, 748-1346, 754-995, 754-1108, 761-1310, 772-912, 772-1020, 773-892, 773-921, 778-1115, 778-1117, 780-887, 781-1241, 784-1035, 785-1097, 789-993, 789-1263, 791-982, 791-995, 791-1011, 791-1023, 791-1041, 792-994, 798-1379, 803-1065, 803-1107, 805-1202, 807-1113, 808-1070, 808-1101, 809-1099, 819-1193, 819-1261, 820-987, 825-1264, 834-955, 834-1086, 834-1250, 834-1265, 835-1087, 835-1330, 837-1310, 840-1017, 840-1133, 844-1105, 846-1094, 846-1126, 853-1262, 856-1099, 856-1111, 870-1125, 870-1148, 871-1011, 873-1312, 876-1317, 894-1157, 894-1319, 896-1117, 906-1214, 909-1315, 911-1160, 917-1157, 918-1191, 921-1155, 927-1355, 929-1212, 938-1212, 949-1186, 950-1207, 950-1236, 950-1249, 951-1188, 956-1224,
	956-1244, 956-1246, 958-1175, 958-1217, 958-1236, 958-1250, 961-1220, 964-1208, 965-1200, 969-1225, 969-1231, 976-1126, 976-1211, 976-1237, 979-1238, 991-1232, 993-1184, 994-1237, 995-1365, 996-1262, 997-1290, 998-1176, 998-1225, 999-1250, 1005-1371, 1007-1321, 1010-1245, 1010-1254, 1010-1280, 1011-1172, 1011-1220, 1011-1254, 1012-1280, 1012-1477, 1012-1488, 1020-1223, 1020-1294, 1023-1209, 1026-1331, 1026-1335, 1026-1443, 1028-1209, 1031-1157, 1031-1316, 1031-1322, 1031-1327, 1031-1328, 1031-1338, 1031-1341, 1031-1349, 1032-1333, 1034-1232, 1040-1289, 1048-1305, 1048-1314, 1048-1326, 1056-1282, 1056-1337, 1058-1296, 1058-1338, 1059-1312, 1062-1334, 1063-1336, 1068-1194, 1068-1308, 1068-1309, 1069-1261, 1069-1331, 1070-1311, 1075-1327, 1076-1221, 1076-1280, 1085-1203, 1086-1252, 1086-1347, 1088-1334, 1095-1286, 1108-1308, 1115-1240, 1115-1346, 1115-1362, 1128-1335, 1133-1218, 1136-1271, 1148-1323, 1154-1282, 1176-1303, 1201-1285, 1215-1371
48/7517764CB1 2320	1-284, 1-2320, 117-343, 117-344, 117-345, 117-348, 117-350, 117-365, 117-369, 118-369, 122-366, 124-359, 132-369, 133-363, 136-359, 153-358, 162-357, 162-362, 242-545, 311-504, 349-1027, 407-671, 678-720, 678-1230, 704-968, 706-1286, 707-980, 711-1342, 714-955, 714-1176, 725-982, 725-1029, 727-941, 730-961, 730-985, 736-1411, 744-944, 744-966, 748-1510, 750-1258, 768-994, 775-1092, 775-1095, 784-1000, 787-1021, 800-1436, 808-1085, 818-1065, 818-1434, 832-1439, 844-1399, 851-1439, 864-1112, 864-1128, 865-1108, 865-1122, 867-1415, 868-1070, 879-1369, 890-1512, 892-1154, 892-1173, 892-1186, 896-1510, 900-1117, 910-1511, 914-1612, 915-1461, 916-1375, 919-1182, 923-1136, 953-1443, 955-1367, 958-1076, 959-1229, 962-1770, 983-1615, 1016-1285, 1026-1493, 1029-1490, 1050-1516, 1051-1481, 1057-1211, 1058-1327, 1073-1270, 1073-1319, 1074-1348, 1078-1330, 1083-1754, 1091-1260, 1091-1659, 1099-1690, 1104-1382, 1109-1367, 1110-1343, 1112-1734, 1114-1389, 1114-1392, 1114-1401, 1115-1550, 1115-1553, 1115-1873, 1117-1407,

Table 4

Polynucleotide SEQ ID NO:/ Incyte ID/ Sequence Length	Sequence Fragments
	1119-1230, 1121-1585, 1130-1374, 1132-1367, 1132-1419, 1132-1818, 1140-1396, 1140-1402, 1140-1533, 1142-1403, 1147-1416, 1147-1449, 1148-1367, 1160-1424, 1172-1442, 1192-1480, 1196-1483, 1202-1632, 1202-1756, 1214-1944, 1215-1518, 1220-1702, 1225-1746, 1226-1515, 1227-1423, 1230-1495, 1233-1839, 1240-1492, 1240-1707, 1244-1499, 1247-1705, 1249-1862, 1260-1857, 1262-1492, 1262-1519, 1263-1542, 1264-1536, 1291-1579, 1292-1513, 1292-1571, 1297-1550, 1308-1612, 1320-1840, 1325-1603, 1329-1600, 1329-1603, 1331-1623, 1331-1625, 1340-1565, 1342-1577, 1342-1584, 1353-1630, 1353-1637, 1366-1510, 1368-1948, 1376-1494, 1381-1635, 1386-1660, 1394-1658, 1394-1930, 1405-2022, 1410-2229, 1416-1693, 1421-1685, 1424-1877, 1426-1662, 1426-1700, 1440-1523, 1440-1725, 1442-1741, 1442-1891, 1443-1803, 1456-1933, 1460-1937, 1463-1838, 1472-1794, 1481-1945, 1488-1681, 1488-1811, 1488-1823, 1488-1838, 1489-1945, 1491-1945, 1496-1947, 1503-1924, 1504-1945, 1506-1753, 1508-1815, 1508-1948, 1510-1942, 1511-1945, 1513-2021, 1527-1821,
	1535-1945, 1536-1674, 1537-1791, 1538-1783, 1538-1881, 1538-1887, 1545-1836, 1546-1875, 1547-1824, 1555-1948, 1555-2201, 1556-1936, 1556-2130, 1567-1942, 1570-1825, 1570-2048, 1579-1812, 1579-1837, 1579-2101, 1580-1945, 1585-1836, 1590-1838, 1600-1845, 1603-2317, 1604-2288, 1613-2226, 1618-2161, 1620-1944, 1623-2288, 1638-1948, 1641-1944, 1655-1914, 1655-2148, 1657-1940, 1659-1944, 1659-2270, 1660-1942, 1660-2288, 1661-1883, 1661-1909, 1661-1915, 1661-1948, 1668-1836, 1668-2053, 1676-1948, 1680-1943, 1689-1962, 1696-1942, 1696-2010, 1698-1923, 1700-1938, 1701-2206, 1703-1943, 1705-1978, 1705-1993, 1707-1948, 1710-1945, 1713-1973, 1714-1958, 1714-2021, 1714-2053, 1723-1948, 1737-2316, 1739-1945, 1740-2277, 1742-2257, 1742-2282, 1745-1889, 1752-2292, 1760-1945, 1761-2293, 1764-2263, 1780-2037, 1780-2047, 1790-2075, 1792-2022, 1794-2038, 1795-2178, 1820-2293, 1835-2084, 1842-2129, 1846-2085, 1858-1948, 1859-2131, 1867-1948, 1878-2320, 1889-2257, 1898-2320, 1904-2320, 1906-2320, 1913-2320, 1914-2316, 1918-2320,
	1919-2129, 1920-2320, 1934-2210, 1949-2172, 1949-2274, 1956-2072, 1956-2198, 1956-2206, 1958-2210, 1963-2320, 1967-2320, 1968-2214, 1971-2320, 1988-2320, 1991-2320, 1992-2296, 1995-2320, 2002-2316, 2016-2320, 2017-2292, 2021-2277, 2021-2315, 2021-2320, 2028-2207, 2031-2235, 2061-2320, 2064-2309, 2079-2320, 2136-2320, 2151-2320, 2153-2320, 2175-2320
49/7517774CB1 2266	1-759, 1-2209, 721-1504, 743-1504, 1430-2266, 1437-2207
50/7518133CB1 1397	1-215, 1-567, 1-607, 1-691, 1-787, 2-1396, 79-787, 524-1254, 524-1396, 667-1397, 694-1397, 779-1397
51/7520147CB1 906	1-903, 1-904, 2-905, 64-906, 80-906, 285-906

Table 4

Polynucleotide SEQ ID NO:/ Incyte ID/ Sequence Length	Sequence Fragments
52/7520276CB1 1326	1-895, 450-1326, 511-1326, 533-1325, 543-1326
53/7520808CB1 1090	1-870, 2-1089, 213-1090
54/7520821CB1 776	1-360, 1-726, 7-81, 7-776, 123-776, 175-198, 299-775, 300-775
55/7520839CB1 549	1-549, 4-549, 23-549, 31-549, 49-549, 62-137, 131-549, 356-544
56/7520891CB1 623	1-618, 1-621, 1-623
57/7514645CB1 1751	1-1751, 484-1185, 484-1228, 484-1237
58/7517776CB1 3010	1-841, 1-3009, 2-818, 585-3009, 671-1543, 671-1549, 672-1313, 674-1349, 676-1440, 1358-2227, 1358-2233, 1359-2222, 1359-2237, 1385-2221, 1442-2230, 1442-2251, 1443-2220, 1443-2228, 2082-3010, 2106-3010, 2146-3010, 2183-3010
59/7517783CB1 3242	1-826, 1-838, 1-3227, 5-826, 675-1471, 679-1558, 1425-3227, 1432-2295, 1460-2181, 1540-2482, 1573-2459, 1574-2434, 1758-2490, 2337-3242, 2377-3242, 2389-3242
60/7522607CB1 1360	1-666, 572-1360
61/7521142CB1 1015	1-880, 2-803, 21-794, 137-1015
62/7521689CB1 1489	1-781, 6-1489, 649-1489
63/2878775CB1 3871	1-364, 2-364, 12-598, 12-3335, 803-1367, 935-1532, 935-1569, 974-1596, 1317-1569, 1317-1587, 1317-1600, 1317-1771, 1317-1812, 1542-1939, 1596-2180, 1802-2478, 1891-2354, 1911-2159, 1922-2570, 1954-2197, 1965-2210, 1965-2506, 2250-2505, 2294-2552, 2305-2479, 2426-2559, 2426-2764, 2488-2751, 2509-3144, 2604-2840, 2604-2861, 2680-3144, 2714-2946, 2783-3084, 2847-3077, 2849-3093, 2849-3136, 2849-3231, 2923-3083, 2938-3650, 2991-3494, 3020-3270, 3020-3557, 3056-3606, 3059-3446, 3063-3624, 3091-3871, 3112-3314, 3151-3744, 3155-3845, 3231-3525, 3264-3862, 3282-3825, 3293-3754, 3307-3519, 3358-3865, 3361-3849, 3366-3862, 3381-3579, 3381-3845, 3381-3852, 3381-3855, 3409-3847, 3409-3856, 3419-3857, 3434-3856, 3440-3862, 3480-3871, 3530-3867, 3580-3809

Table 4

Polynucleotide SEQ ID NO:/ Incyte ID/ Sequence Length	Sequence Fragments
64/7521207CB1 270	1-270, 4-270
65/7521283CB1 318	1-318
66/7522210CB1 1216	1-890, 454-1216
67/7519488CB1 1306	1-729, 1-779, 1-805, 1-856, 1-1306, 350-1306, 466-1306, 547-1306, 587-1306, 815-1306, 869-1306
68/7519965CB1 1321	1-410, 1-422, 1-701, 1-818, 2-1320, 407-1321, 421-1321, 449-1321, 485-1321, 535-1321, 694-1321, 695-1321, 783-1321, 828-1321, 831-1321
69/7519985CB1 676	1-302, 1-676, 2-675, 379-676
70/7520002CB1 1014	1-670, 1-770, 2-739, 48-1014
71/7520014CB1 991	1-965, 10-746, 10-924, 11-990, 195-991, 203-991, 236-991, 242-991, 470-991
72/7520039CB1 545	1-545, 2-544
73/7520053CB1 831	1-831, 2-830, 88-831, 221-831, 535-831
74/7523262CB1 888	1-888, 2-566, 5-888
75/7523270CB1 795	1-710, 1-795, 2-789, 501-795
76/7523287CB1 1174	1-862, 2-1173, 381-1174, 695-1174
77/7521825CB1 1159	1-467, 1-538, 1-545, 1-761, 1-903, 1-910, 2-1158, 239-1159, 290-1159, 539-1159

Table 4

Polynucleotide SEQ ID NO:/ Incyte ID/ Sequence Length	Sequence Fragments
78/7521844CB1 813	1-541, 1-615, 1-689, 1-813, 2-510, 2-812
79/7521864CB1 503	1-190, 1-374, 1-502, 1-503, 2-221, 2-501, 104-226, 104-256, 104-260, 104-270, 104-329, 104-502, 104-503, 109-329, 109-503, 130-389, 133-389, 152-389, 167-253, 167-274, 167-389, 172-389, 193-379, 196-379, 213-379, 230-316, 230-379, 235-379
80/7522020CB1 805	1-805, 2-804, 4-805
81/758410CB1 3140	1-483, 1-665, 322-801, 556-1402, 598-720, 598-801, 599-666, 666-999, 691-901, 750-801, 826-1572, 998-1048, 998-1051, 998-1062, 1045-1599, 1046-1610, 1068-1402, 1068-1602, 1068-1625, 1070-1513, 1070-1625, 1246-1478, 1293-1978, 1302-1442, 1400-1643, 1482-1785, 1671-2235, 1672-1934, 1755-2026, 1755-2197, 1760-1916, 1787-2175, 1937-2174, 1937-2405, 1974-2248, 1997-2287, 2005-2300, 2005-2706, 2036-2279, 2067-2706, 2080-2261, 2167-2612, 2209-2595, 2251-2706, 2288-2706, 2307-2706, 2341-2495, 2366-2706, 2436-2706, 2523-2706, 2535-2706, 2582-2706, 2641-2706, 2649-2706, 2661-2706, 2700-2806, 2700-2823, 2700-3105, 2700-3115, 2700-3140, 2702-3115, 2734-3115, 2746-3115, 2793-3115, 2852-3115, 2860-3092, 2872-3115, 2911-3115, 2935-3115, 2968-3115, 2990-3115, 2996-3115, 3003-3115, 3071-3115, 3082-3115
82/7520759CB1 1119	1-621, 265-1119, 434-1118, 667-1119, 742-1118, 860-1119
83/7522915CB1 1319	1-721, 1-780, 2-690, 2-1319, 7-780, 611-1319, 635-1319, 707-1319
84/7522936CB1 1212	1-744, 2-1211, 405-1212

Table 5

Polynucleotide SEQ ID NO:	Incyte Project ID:	Representative Library
43	7513225CB1	BRAITUE01
44	7513288CB1	HNT2NOT01
47	7513298CB1	FIBRTXS07
48	7517764CB1	FIBPFEN06
63	2878775CB1	BRAITUT08
81	758410CB1	BRAITUT02

Table 6

Library	Vector	Library Description
BRAITUE01	PCDNA2.1	This 5' biased random primed library was constructed using RNA isolated from brain meningioma tissue removed from a 35-year-old Caucasian female during excision of a cerebral meningeal lesion. Pathology indicated a benign neoplasm in the right cerebellopontine angle of the brain. The patient presented with headache and deficiency anemia. Patient history included hypothyroidism. Patient medications included Synthroid. Family history included a myocardial infarction in the father, breast cancer in the mother, alcohol abuse in the grandparent(s), and drug-induced mental disorder in the sibling(s).
BRAITUT02	PSPORT1	Library was constructed using RNA isolated from brain tumor tissue removed from the frontal lobe of a 58-year-old Caucasian male during excision of a cerebral meningeal lesion. Pathology indicated a grade 2 metastatic hypernephroma. Patient history included a grade 2 renal cell carcinoma, insomnia, and chronic airway obstruction. Family history included a malignant neoplasm of the kidney.
BRAITUT08	pINCY	Library was constructed using RNA isolated from brain tumor tissue removed from the left frontal lobe of a 47-year-old Caucasian male during excision of cerebral meningeal tissue. Pathology indicated grade 4 fibrillary astrocytoma with focal tumoral radionecrosis. Patient history included cerebrovascular disease, deficiency anemia, hyperlipidemia, epilepsy, and tobacco use. Family history included cerebrovascular disease and a malignant prostate neoplasm.
FIBPFEN06	pINCY	The normalized prostate stromal fibroblast tissue libraries were constructed from 1.56 million independent clones from a prostate fibroblast library. Starting RNA was made from fibroblasts of prostate stroma removed from a male fetus, who died after 26 weeks' gestation. The libraries were normalized in two rounds using conditions adapted from Soares et al., PNAS (1994) 91:9228 and Bonaldo et al., Genome Research (1996) 6:791, except that a significantly longer (48-hours/round)reannealing hybridization was used. The library was then linearized and recircularized to select for insert containing clones as follows: plasmid DNA was prepped from approximately 1 million clones from the normalized prostate stromal fibroblast tissue libraries following soft agar transformation.
FIBRTXS07	pINCY	This subtracted library was constructed using 1.3 million clones from a dermal fibroblast library and was subjected to two rounds of subtraction hybridization with 2.8 million clones from an untreated dermal fibroblast tissue library. The starting library for subtraction was constructed using RNA isolated from treated dermal fibroblast tissue removed from the breast of a 31-year-old Caucasian female. The cells were treated with 9CIS retinoic acid. The hybridization probe for subtraction was derived from a similarly constructed library from RNA isolated from untreated dermal fibroblast tissue from the same donor. Subtractive hybridization conditions were based on the methodologies of Swaroop et al., NAR (1991) 19:1954 and Bonaldo, et al., Genome Research (1996) 6:791.

Table 6

Library	Vector	Library Description
HNT2NOT01	PBLUESCRIPT	Library was constructed at Stratagene (STR937230), using RNA isolated from the hNT2 cell line (derived from a human teratocarcinoma that exhibited properties characteristic of a committed neuronal precursor).

Table 7

Program	Description	Reference	Parameter Threshold
ABI FACTURA	A program that removes vector sequences and masks ambiguous bases in nucleic acid sequences.	Applied Biosystems, Foster City, CA.	
ABI/PARACEL FDF	A Fast Data Finder useful in comparing and annotating amino acid or nucleic acid sequences.	Applied Biosystems, Foster City, CA; Paracel Inc., Pasadena, CA.	Mismatch <50%
ABI AutoAssembler	A program that assembles nucleic acid sequences.	Applied Biosystems, Foster City, CA.	
BLAST	A Basic Local Alignment Search Tool useful in sequence similarity search for amino acid and nucleic acid sequences. BLAST includes five functions: blastp, blastn, blastx, tblastn, and tblastx.	Altschul, S.F. et al. (1990) J. Mol. Biol. 215:403-410; Altschul, S.F. et al. (1997) Nucleic Acids Res. 25:3389-3402.	*ESTs:* Probability value=1.0E-8 or less *Full Length sequences*: Probability value= 1.0E-10 or less
FASTA	A Pearson and Lipman algorithm that searches for similarity between a query sequence and a group of sequences of the same type. FASTA comprises as least five functions: fasta, tfasta, fastx, tfastx, and ssearch.	Pearson, W.R. and D.J. Lipman (1988) Proc. Natl. Acad Sci. USA 85:2444-2448; Pearson, W.R. (1990) Methods Enzymol. 183:63-98; and Smith, T.F. and M.S. Waterman (1981) Adv. Appl. Math. 2:482-489.	*ESTs:* fasta E value=1.06E-6 *Assembled ESTs*: fasta Identity= 95% or greater and Match length=200 bases or greater; fastx E value=1.0E-8 or less *Full Length sequences:* fastx score=100 or greater
BLIMPS	A BLocks IMProved Searcher that matches a sequence against those in BLOCKS, PRINTS, DOMO, PRODOM, and PFAM databases to search for gene families, sequence homology, and structural fingerprint regions.	Henikoff, S. and J.G. Henikoff (1991) Nucleic Acids Res. 19:6565-6572; Henikoff, J.G. & S. Henikoff (1996) Methods Enzymol. 266:88-105; and Attwood, T.K. et al. (1997) J. Chem. Inf. Comput. Sci. 37:417-424.	Probability value=1.0E-3 or less
HMMER	An algorithm for searching a query sequence against hidden Markov model (HMM)-based databases of protein family consensus sequences, such as PFAM, INCY, SMART, and TIGRFAM.	Krogh, A. et al. (1994) J. Mol. Biol. 235:1501-1531; Sonnhammer, E.L.L. et al. (1988) Nucleic Acids Res. 26:320-322; Durbin, R. et al. (1998) Our World View, in a Nutshell, Cambridge Univ. Press, p. 1-350	*PFAM, INCY, SMART, or TIGRFAM hits:* Probability value=1.0E-3 or less *Signal peptide hits:* Score= 0 or greater

Table 7 (cont.)

Program	Description	Reference	Parameter Threshold
ProfileScan	An algorithm that searches for structural and sequence motifs in protein sequences that match sequence patterns defined in Prosite.	Gribskov, M. et al. (1988) CABIOS 4:61-66; Gribskov, M. et al. (1989) Methods Enzymol. 183:146-159; Bairoch, A. et al. (1997) Nucleic Acids Res. 25:217-221.	Normalized quality score≥GCG-specified "HIGH" value for that particular Prosite motif. Generally, score=1.4-2.1.
Phred	A base-calling algorithm that examines automated sequencer traces with high sensitivity and probability.	Ewing, B. et al. (1998) Genome Res. 8:175-185; Ewing, B. and P. Green (1998) Genome Res. 8:186-194.	
Phrap	A Phils Revised Assembly Program including SWAT and CrossMatch, programs based on efficient implementation of the Smith-Waterman algorithm, useful in searching sequence homology and assembling DNA sequences.	Smith, T.F. and M.S. Waterman (1981) Adv. Appl. Math. 2:482-489; Smith, T.F. and M.S. Waterman (1981) J. Mol. Biol. 147:195-197; and Green, P., University of Washington, Seattle, WA.	Score=120 or greater; Match length=56 or greater
Consed	A graphical tool for viewing and editing Phrap assemblies.	Gordon, D. et al. (1998) Genome Res. 8:195-202.	
SPScan	A weight matrix analysis program that scans protein sequences for the presence of secretory signal peptides.	Nielson, H. et al. (1997) Protein Engineering 10:1-6; Claverie, J.M. and S. Audic (1997) CABIOS 12:431-439.	Score=3.5 or greater
TMAP	A program that uses weight matrices to delineate transmembrane segments on protein sequences and determine orientation.	Persson, B. and P. Argos (1994) J. Mol. Biol. 237:182-192; Persson, B. and P. Argos (1996) Protein Sci. 5:363-371.	
TMHMMER	A program that uses a hidden Markov model (HMM) to delineate transmembrane segments on protein sequences and determine orientation.	Sonnhammer, E.L. et al. (1998) Proc. Sixth Intl. Conf. on Intelligent Systems for Mol. Biol., Glasgow et al., eds., The Am. Assoc. for Artificial Intelligence Press, Menlo Park, CA, pp. 175-182.	
Motifs	A program that searches amino acid sequences for patterns that matched those defined in Prosite.	Bairoch, A. et al. (1997) Nucleic Acids Res. 25:217-221; Wisconsin Package Program Manual, version 9, page M51-59, Genetics Computer Group, Madison, WI.	

Table 8

SEQ ID NO:	PID	EST ID	SNP ID	EST SNP	CB1 SNP	EST Allele	Allele 1	Allele 2	Amino Acid	Caucasian Allele 1 frequency	African Allele 1 frequency	Asian Allele 1 frequency	Hispanic Allele 1 frequency
43	7513225	1395701H1	SNP00046418	70	3720	A	A	G	N1222	n/d	n/d	n/d	n/d
43	7513225	1404966T6	SNP00046416	248	4365	G	G	A	noncoding	n/a	n/a	n/a	n/a
43	7513225	1443205H1	SNP00046417	250	4144	C	C	T	noncoding	n/a	n/a	n/a	n/a
43	7513225	1807119H1	SNP00046421	233	2260	T	C	T	D735	n/a	n/a	n/a	n/a
43	7513225	2256848T6	SNP00046416	249	4363	G	G	A	noncoding	n/a	n/a	n/a	n/a
43	7513225	4581582H1	SNP00046420	154	2917	G	G	A	P954	n/a	n/a	n/a	n/a
43	7513225	4751423F6	SNP00069222	389	3016	G	A	G	R987	n/d	n/a	n/a	n/a
43	7513225	6472082H1	SNP00069223	330	1633	G	A	G	E526	n/a	n/a	n/a	n/a
43	7513225	6472082H1	SNP00124671	108	1413	G	G	A	G453	0.18	0.48	0.49	0.42
44	7513288	1352889H1	SNP00130515	99	4278	T	T	C	D1383	n/a	n/a	n/a	n/a
44	7513288	1452458H1	SNP00071405	140	5981	C	C	A	A1951	n/a	n/a	n/a	n/a
44	7513288	1453633F6	SNP00061745	77	825	A	G	A	V232	0.41	0.47	0.26	0.49
44	7513288	4704817H1	SNP00130516	49	4774	T	T	C	W1549	n/a	n/a	n/a	n/a
44	7513288	6298026H1	SNP00124562	88	3739	A	A	G	T1204	n/a	n/a	n/a	n/a
44	7513288	6298026H1	SNP00124563	163	3814	A	A	G	T1229	n/a	n/a	n/a	n/a
44	7513288	6326804H1	SNP00021486	53	5101	A	G	A	R1658	n/a	n/a	n/a	n/a
44	7513288	7695619H1	SNP00056540	111	4703	A	G	A	D1525	n/a	n/a	n/a	n/a
45	7513607	2989187H1	SNP00014113	166	1373	A	G	A	noncoding	0.66	n/a	n/a	n/a
46	7513991	148825T6	SNP00036263	13	781	C	C	A	noncoding	n/a	n/a	n/a	n/a
46	7513991	1703775H1	SNP00019459	38	38	G	G	A	noncoding	n/a	n/a	n/a	n/a
46	7513991	1703775T6	SNP00093433	24	745	T	T	C	noncoding	n/a	n/a	n/a	n/a
46	7513991	2477016H1	SNP00093433	213	744	T	T	C	noncoding	n/a	n/a	n/a	n/a
46	7513991	2729192H1	SNP00036263	172	774	C	C	A	noncoding	n/a	n/a	n/a	n/a
46	7513991	3137269H1	SNP00114333	79	401	G	G	A	P75	n/a	n/a	n/a	n/a
46	7513991	8625909J1	SNP00114333	401	420	G	G	A	E82	n/a	n/a	n/a	n/a
47	7513298	1237992H1	SNP00092742	206	1293	C	G	C	noncoding	n/a	n/a	n/a	n/a
47	7513298	1237992H1	SNP00093120	66	1153	G	G	A	noncoding	0.94	n/a	n/a	n/a
47	7513298	1237992H1	SNP00128327	92	1179	T	T	C	noncoding	n/a	n/a	n/a	n/a

Table 8

SEQ ID NO:	PID	EST ID	SNP ID	EST SNP	CB1 SNP	EST Allele	Allele 1	Allele 2	Amino Acid	Caucasian Allele 1 frequency	African Allele 1 frequency	Asian Allele 1 frequency	Hispanic Allele 1 frequency
47	7513298	1251132H1	SNP00128325	99	596	C	C	T	P165	n/a	n/a	n/a	n/a
47	7513298	1298507H1	SNP00128326	176	1010	C	C	A	N303	n/a	n/a	n/a	n/a
47	7513298	1319020F6	SNP00055512	98	363	C	C	T	L88	n/d	n/a	n/a	n/a
47	7513298	1319020F6	SNP00115384	205	470	C	C	G	H123	n/a	n/a	n/a	n/a
47	7513298	1342950H1	SNP00001611	98	108	A	C	A	S3	n/a	n/a	n/a	n/a
47	7513298	1342950H1	SNP00055511	88	98	C	C	T	noncoding	n/a	n/a	n/a	n/a
47	7513298	1353887H1	SNP00128329	182	1434	C	C	A	noncoding	n/a	n/a	n/a	n/a
47	7513298	1421034F6	SNP00128324	302	236	A	A	G	S45	n/a	n/a	n/a	n/a
47	7513298	3770642H1	SNP00092746	239	1327	G	G	C	noncoding	n/a	n/a	n/a	n/a
47	7513298	3779622H1	SNP00128328	86	1314	C	C	T	noncoding	n/a	n/a	n/a	n/a
47	7513298	4534891T1	SNP00128329	310	1468	C	C	A	noncoding	n/a	n/a	n/a	n/a
47	7513298	6362171H1	SNP00001612	434	195	C	A	C	L32	0.68	n/a	n/a	n/a
47	7513298	7644907J1	SNP00115384	103	475	C	C	G	S125	n/a	n/a	n/a	n/a
48	7517764	1002205H1	SNP00004748	37	148	G	G	A	noncoding	n/a	n/a	n/a	n/a
48	7517764	1253069H1	SNP00154954	208	1317	C	C	T	noncoding	n/a	n/a	n/a	n/a
48	7517764	1442033H1	SNP00154955	167	1428	T	T	C	noncoding	n/a	n/a	n/a	n/a
48	7517764	1569549H1	SNP00004749	169	796	C	C	T	noncoding	n/a	n/a	n/a	n/a
48	7517764	1921051H1	SNP00054432	101	847	C	C	T	noncoding	n/a	n/a	n/a	n/a
48	7517764	2098028H1	SNP00025640	117	1016	G	G	T	noncoding	n/a	n/a	n/a	n/a
48	7517764	2207032H1	SNP00025641	181	1139	T	T	C	noncoding	n/a	n/a	n/a	n/a
48	7517764	5196459H1	SNP00119380	94	739	C	C	A	noncoding	n/d	n/d	n/d	n/d
48	7517764	7735718J1	SNP00025641	372	1138	T	T	C	noncoding	n/a	n/a	n/a	n/a
50	7518133	1456901H1	SNP00027692	199	713	C	C	G	P231	n/a	n/a	n/a	n/a
50	7518133	1794980H1	SNP00052100	268	953	G	G	C	G311	n/d	n/a	n/a	n/a
50	7518133	3403559H1	SNP00052099	208	465	G	G	A	R148	n/a	n/a	n/a	n/a
50	7518133	3403559H1	SNP00098207	90	347	C	G	C	A109	n/a	n/a	n/a	n/a
50	7518133	3752762F6	SNP00027692	464	714	C	C	G	P231	n/a	n/a	n/a	n/a
50	7518133	3752762F6	SNP00052099	216	466	G	G	A	V149	n/a	n/a	n/a	n/a

Table 8

SEQ ID NO:	PID	EST ID	SNP ID	EST SNP	CB1 SNP	EST Allele	Allele 1	Allele 2	Amino Acid	Caucasian Allele 1 frequency	African Allele 1 frequency	Asian Allele 1 frequency	Hispanic Allele 1 frequency
50	7518133	3752762F6	SNP00098207	98	348	C	G	C	G109	n/a	n/a	n/a	n/a
51	7520147	8094509H1	SNP00134923	246	172	A	A	G	N52	n/a	n/a	n/a	n/a
52	7520276	1273535H1	SNP00130807	48	1207	C	C	T	noncoding	n/a	n/a	n/a	n/a
52	7520276	1362236H1	SNP00128942	127	552	C	C	T	L178	n/a	n/a	n/a	n/a
52	7520276	1444387H1	SNP00098815	168	457	C	C	T	P146	n/d	n/d	0.98	n/d
52	7520276	1897461H2	SNP00092733	190	1151	C	C	T	noncoding	n/d	n/d	n/d	n/d
52	7520276	1897461H2	SNP00149433	26	983	A	A	G	noncoding	n/a	n/a	n/a	n/a
53	7520808	1894549H1	SNP00112590	190	274	G	G	A	G82	0.89	n/a	0.75	0.99
53	7520808	2198035H1	SNP00020738	173	297	C	C	G	V89	0.96	0.78	n/d	0.9
53	7520808	2200628H1	SNP00053538	194	648	C	C	T	F206	n/d	n/a	n/a	n/a
53	7520808	3642506F6	SNP00020738	381	296	C	C	G	A89	0.96	0.78	n/d	0.9
53	7520808	7613106H1	SNP00112590	169	277	A	G	A	T83	0.89	n/a	0.75	0.99
53	7520808	7615381H1	SNP00143739	396	856	T	C	T	stop276	n/a	n/a	n/a	n/a
54	7520821	1375496H1	SNP00011005	89	665	A	G	A	noncoding	0.81	0.79	0.38	0.6
54	7520821	1406791T6	SNP00011005	393	656	G	G	A	noncoding	0.81	0.79	0.38	0.6
54	7520821	1466811T6	SNP00011005	388	672	G	G	A	noncoding	0.81	0.79	0.38	0.6
54	7520821	2815567H1	SNP00011004	132	578	T	C	T	noncoding	n/a	n/a	n/a	n/a
54	7520821	4075873H1	SNP00050280	188	458	C	C	T	noncoding	0.94	n/a	n/a	n/a
58	7517776	5968989H1	SNP00113454	69	2421	G	G	C	noncoding	0.99	n/a	n/a	n/a
59	7517783	5968989H1	SNP00113454	69	2653	G	G	C	V865	0.99	n/a	n/a	n/a
61	7521142	2988106H1	SNP00058530	27	470	T	C	T	noncoding	0.58	0.73	0.81	0.66
62	7521689	1351837T6	SNP00043567	533	1168	C	T	C	noncoding	0.63	0.62	0.84	n/a
62	7521689	2179888F6	SNP00043566	390	349	C	C	T	D59	n/a	n/a	n/a	n/a
62	7521689	2555648H1	SNP00043567	43	1166	C	T	C	noncoding	0.63	0.62	0.84	n/a
62	7521689	2846807H1	SNP00043565	173	59	A	G	A	noncoding	0.74	0.66	0.84	0.82
63	2878775	2656707F6	SNP00105335	341	3360	T	T	C	noncoding	n/a	n/a	n/a	n/a
63	2878775	2878775F6	SNP00105334	63	1379	G	A	G	V416	n/d	n/d	0.01	0.01
66	7522210	1421702H1	SNP00098613	62	8	C	C	T	noncoding	0.8	n/a	n/a	n/a

Table 8

SEQ ID NO:	PID	EST ID	SNP ID	EST SNP	CB1 SNP	EST Allele	Allele 1	Allele 2	Amino Acid	Caucasian Allele 1 frequency	African Allele 1 frequency	Asian Allele 1 frequency	Hispanic Allele 1 frequency
66	7522210	1503756F6	SNP00047003	140	869	C	T	C	T280	0.63	n/a	n/a	n/a
66	7522210	2561551H1	SNP00035743	73	769	T	T	C	C247	n/a	n/a	n/a	n/a
66	7522210	2989336H1	SNP00008542	224	1051	A	G	A	N341	n/a	n/a	n/a	n/a
66	7522210	6332517H1	SNP00058787	373	1154	G	C	G	noncoding	n/a	n/a	n/a	n/a
66	7522210	7619433H1	SNP00058786	286	376	A	A	G	I116	n/a	n/a	n/a	n/a
66	7522210	7696491J1	SNP00035743	21	764	T	T	C	M245	n/a	n/a	n/a	n/a
66	7522210	998626R6	SNP00038825	304	100	A	A	G	T24	n/d	n/a	n/a	n/a
67	7519488	663873H1	SNP00097527	80	1125	G	T	G	noncoding	n/d	n/a	n/a	n/a
68	7519965	663873H1	SNP00097527	80	1139	G	T	G	noncoding	n/d	n/a	n/a	n/a
69	7519985	2651464F6	SNP00152518	287	605	A	A	G	noncoding	n/a	n/a	n/a	n/a
69	7519985	2651464T6	SNP00152518	71	599	A	A	G	noncoding	n/a	n/a	n/a	n/a
69	7519985	4825157H1	SNP00152518	52	592	A	A	G	noncoding	n/a	n/a	n/a	n/a
71	7520014	8094509H1	SNP00134923	246	181	A	A	G	N52	n/a	n/a	n/a	n/a
71	7520014	8094509H1	SNP00134925	476	411	T	C	T	C128	n/a	n/a	n/a	n/a
71	7520014	8096114H1	SNP00134924	436	369	G	G	A	G114	n/a	n/a	n/a	n/a
72	7520039	1357120H1	SNP00136436	117	262	T	T	C	S81	0.99	n/a	n/a	n/a
72	7520039	2110750H1	SNP00053655	125	522	G	G	A	R167	n/a	n/a	n/a	n/a
72	7520039	2160613H1	SNP00050861	88	19	G	G	T	noncoding	0.99	n/a	n/a	n/a
73	7520053	2651464F6	SNP00152518	287	761	A	A	G	noncoding	n/a	n/a	n/a	n/a
73	7520053	2651464T6	SNP00152518	71	755	A	A	G	noncoding	n/a	n/a	n/a	n/a
73	7520053	4825157H1	SNP00152518	52	748	A	A	G	noncoding	n/a	n/a	n/a	n/a
75	7523270	1274681F1	SNP00000474	382	585	C	C	T	I192	n/a	n/a	n/a	n/a
75	7523270	1274681F1	SNP00114404	460	663	T	T	C	G218	0.64	n/a	n/a	n/a
75	7523270	1454536F6	SNP00000475	14	741	C	C	T	P244	n/a	n/a	n/a	n/a
75	7523270	7732768J2	SNP00000474	418	587	C	C	T	A193	n/a	n/a	n/a	n/a
75	7523270	7732768J2	SNP00000475	262	743	C	C	T	A245	n/a	n/a	n/a	n/a
75	7523270	7732768J2	SNP00114404	340	665	C	T	C	T219	0.64	n/a	n/a	n/a
76	7523287	1356557H1	SNP00074692	193	711	G	G	A	noncoding	n/a	n/a	n/a	n/a

Table 8

SEQ ID NO:	PID	EST ID	SNP ID	EST SNP	CB1 SNP	EST Allele	Allele 1	Allele 2	Amino Acid	Caucasian Allele 1 frequency	African Allele 1 frequency	Asian Allele 1 frequency	Hispanic Allele 1 frequency
76	7523287	1503624F6	SNP00074693	217	1156	T	T	G	noncoding	n/a	n/a	n/a	n/a
76	7523287	1725519H1	SNP00033549	89	491	C	C	A	D160	n/d	n/d	n/d	n/d
76	7523287	3487554H1	SNP00074693	217	1153	T	T	G	noncoding	n/a	n/a	n/a	n/a
76	7523287	4079996F7	SNP00123151	176	796	C	C	T	noncoding	n/d	0.94	n/d	0.98
77	7521825	1573175H1	SNP00037503	17	943	C	C	T	Y305	n/a	n/a	n/a	n/a
78	7521844	1423977H1	SNP00040822	148	469	G	A	G	R144	n/a	n/a	n/a	n/a
78	7521844	3142364F6	SNP00040822	395	470	A	A	G	Q144	n/a	n/a	n/a	n/a
79	7521864	1323261H1	SNP00015392	190	379	C	C	G	P125	n/a	n/a	n/a	n/a
80	7522020	1231737R6	SNP00022156	311	307	C	C	T	S102	n/a	n/a	n/a	n/a
80	7522020	1231737T6	SNP00022156	307	306	C	C	T	S102	n/a	n/a	n/a	n/a
81	758410	1441115F6	SNP00000192	111	2045	A	A	G	noncoding	0.81	0.58	0.92	0.88
81	758410	1441115F6	SNP00000193	323	2257	T	T	C	noncoding	0.47	0.11	0.25	0.45
81	758410	1441115R1	SNP00058863	382	2599	G	A	G	noncoding	n/a	n/a	n/a	n/a
81	758410	2578520H1	SNP00014844	76	2113	A	A	G	noncoding	n/a	n/a	n/a	n/a
81	758410	2887987F6	SNP00000192	291	2044	A	A	G	noncoding	0.81	0.58	0.92	0.88
81	758410	3407829H1	SNP00151175	225	2967	T	T	C	noncoding	n/a	n/a	n/a	n/a
81	758410	6008196H1	SNP00058864	35	2927	G	G	A	noncoding	n/a	n/a	n/a	n/a
81	758410	758410R6	SNP00000193	253	2256	C	T	C	noncoding	0.47	0.11	0.25	0.45
81	758410	758410R6	SNP00014844	109	2112	A	A	G	noncoding	n/a	n/a	n/a	n/a
81	758410	758410T6	SNP00058863	665	2207	G	A	G	noncoding	n/a	n/a	n/a	n/a
81	758410	758410T6	SNP00058864	548	2325	A	G	A	noncoding	n/a	n/a	n/a	n/a
81	758410	758410T6	SNP00151175	508	2365	T	T	C	noncoding	n/a	n/a	n/a	n/a
82	7520759	1562752T6	SNP00035104	80	1012	C	C	T	H321	n/a	n/a	n/a	n/a
82	7520759	1562752T6	SNP00071294	251	841	C	C	A	S264	n/d	n/a	n/a	n/a
82	7520759	1880448T6	SNP00035103	242	854	C	C	T	R269	n/a	n/a	n/a	n/a
82	7520759	1880448T6	SNP00071294	148	949	C	C	A	D300	n/a	n/a	n/a	n/a
82	7520759	2464756H1	SNP00035104	49	967	C	C	T	N306	n/a	n/a	n/a	n/a
82	7520759	2757313H1	SNP00071294	30	796	C	C	A	Y249	n/d	n/a	n/a	n/a

Table 8

SEQ ID NO:	PID	EST ID	SNP ID	EST SNP	CB1 SNP	EST Allele	Allele 1	Allele 2	Amino Acid	Caucasian Allele 1 frequency	African Allele 1 frequency	Asian Allele 1 frequency	Hispanic Allele 1 frequency
82	7520759	2850837T6	SNP00035103	309	707	C	C	T	R220	n/a	n/a	n/a	n/a
82	7520759	2850837T6	SNP00071294	216	800	C	C	A	L251	n/d	n/a	n/a	n/a
82	7520759	5606518F6	SNP00035103	70	703	T	C	T	I218	n/a	n/a	n/a	n/a
84	7522936	663873H1	SNP00097527	80	1030	G	T	G	noncoding	n/d	n/a	n/a	n/a
84	7522936	8575391T1	SNP00148818	545	778	G	A	G	G246	n/a	n/a	n/a	n/a